Universitatea "Babeş-Bolyai"
Facultatea de Sociologie şi Asistenţă Socială
Învăţământ la distanţă

Conf. univ. dr. Cornelia Mureşan

STATISTICĂ SOCIALĂ ŞI ANALIZA COMPUTERIZATA A DATELOR

2010

CUPRINS

Informații generale 5
Date de identificare a cursului 5
Condiționări şi cunoştințe prerechizite 5
Descrierea cursului 5
Organizarea temelor în cadrul cursului 5
Formatul si tipul activităților implicate de curs 5
Materiale bibliografice obligatorii 6
Materiale si instrumente necesare pentru curs 6
Calendar al cursului 7
Politica de evaluare si notare 7
Elemente de deontologie academica 7
Studenți cu dizabilități: 7
Strategii de studiu recomandate: 8

Modul 1 9

Introducere în statistică 9
1.1. Obiectul statisticii 9
1.2. Rolul statisticii în cercetarea ştiințifică 10
1.3. Utilizarea statisticii în profesia de asistent social 10
1.4. Dificultăți şi riscuri în însuşirea metodei statistice 11
1.5. Noțiuni şi concepte utilizate în statistică 12
1.6. Variabile statistice şi măsurarea lor 13
1.7. Scale de măsurare a variabilelor statistice 14
Întrebări pentru studiu 17
Exercițiu pentru acasă 1 17

Modul 2 19

Distribuții de frecvențe 19
2.1. Ordonarea datelor brute 19
2.2. Distribuții (tabele) de frecvențe 19
2.3. Tabele de frecvențe pentru variabile cu intervale de grupare 22
2.4. Utilizarea frecvențelor absolute şi relative în comparații 24
Întrebări pentru studiu 26
Exercițiu pentru acasă 2 26

Modul 3 27

Prezentarea grafică a distribuțiilor de frecvențe 27
3.1. Diagrama de bare şi histograma 28
3.2. Diagrama circulară 30
3.3. Poligonul de frecvențe 31
3.4. Diagrama de împrăştiere 32
Întrebări pentru studiu 34
Exercițiu pentru acasă 3 34

Modul 4 35

Indicatori ai tendinței centrale 35
4.1. Modul 35
4.2. Mediana 37
4.3. Media 38
4.4. Modul mediana sau media ? 40
Întrebări pentru studiu 41
Exercițiu pentru acasă 4 42

Modul 5 ..... **43**

Indicatori ai împrăştierii **43**
5.1. Amplitudinea 43
5.2. Abaterea intercuartilă 44
5.3. Abaterea medie 46
5.4. Abaterea standard şi varianța 47
5.5. Algoritmul de calcul al abaterii standard 48
5.6. Utilizarea variantei ca măsură a reducerii gradului de nedeterminare. Indicatorul statistic *eta* 50
Întrebări pentru studiu 53
Exercițiu pentru acasă 5 54

Modul 6 ..... **55**

Distribuția normală **55**
6.1. Proprietăți ale distribuției normale 57
6.2. Valori Z şi distribuția normală standardizată 58
6.3. Utilizarea valorii Z în comparații: un exemplu practic 61
Întrebări pentru studiu 62
Exercitiu pentru acasă 6 63

Modul 7 ..... **64**

Introducere in SPSS **64**
7.1. Procesul de analiză statistică a datelor în SPSS 64
7.2. Procedura Frequencies 69
7.3. Procedura Crosstabs 70
7.4. Procedura Descriptives 71
7.5. Procedura Means 72
7.6. Procedura Explore 73

Modul 8. ..... **76**

Notiuni introductive de statistica inferentiala **76**
8.1. Testarea ipotezelor 76
8.2. Explicatii alternative 76
8.3. Respingerea explicației şansei prin testele statistice 78
8.4. Dovezi suficiente şi semnificație 80
8.5. Relații statistic semnificative şi rezultate substanțiale 81
Întrebări pentru studiu 81
Exercitiu pentru acasă 7 82

Modul 9 ..... **83**

Asocierea **83**
9.1. Ce caută să determine asocierea 83
9.2. Tabele de asociere 83
9.3. Când să nu utilizăm asocierea 89
9.4. Asocierea cu trei sau mai multe variabile 91
9.5. Un exemplu 92
9.6. Folosirea SPSS în analiza asocierii 94
Întrebări pentru studiu 95
Exercițiu pentru acasă 8 95

Modul 10 ..... **96**

Compararea mediilor **96**
10.1. Popularitatea statisticii *t* 96
10.2. Logica statisticii *t* 96
10.3. Calcularea şi interpretarea valorilor lui *t* 97

10.4. Prezentarea statisticii *t* 102
10.5. Când nu se foloseşte testul *t* 102
10.6. Un exemplu 103
10.7. Folosirea SPSS in compararea mediilor 104
Întrebări pentru studiu 106
Exerciţiu pentru acasă 9 106

Modul 11 107

Corelaţia 107
11.1. Conceptul de corelaţie 107
11.2. Coeficientul de corelaţie şi proprietăţile lui fundamentale 109
11.3. Corelaţia cu trei sau mai multe variabile 114
11.4. Un exemplu 114
11.5. Folosirea SPSS in analizele de corelaţie 116
Întrebări pentru studiu 117
Exerciţiu pentru acasă 10 117

***ANEXE* 118**
Bibliografia completa a cursului 118
Scurtă biografie a titularului de curs 118

Informații generale

Date de identificare a cursului

Date de contact ale titularului de curs:

Nume: Conf.univ.dr. Cornelia Mureşan
Birou: Birou 204 sediul Fac. de Sociologie şi asistență socială, bd. 21 Decembrie 1989, nr.128
Telefon: 0264-424674
Fax: 0264-424674
E-mail: cornelia.muresan@socasis.ubbcluj.ro
Consultații: joi, 10-12

Date de identificare curs şi contact tutori:

Numele cursului – Statistică socială şi analiza computerizata a datelor
Codul cursului – AAR0121
Anul, Semestrul – anul 1, sem. 2
Tipul cursului - Obligatoriu
Pagina web a cursului- http://www.asis.ubbcluj.ro
Tutori: Lector dr. Paul-Teodor Hărăguş
tpharagus@ socasis.ubbcluj.ro

Condiționări şi cunoştințe prerechizite

Nu sunt condiționări pentru înscrierea la acest curs. Cunoştințele generale de matematică şi informatică de liceu vor fi foarte utile în rezolvarea lucrărilor de evaluare ce încheie fiecare modul şi, respectiv în promovarea examenului de evaluare finală.

Descrierea cursului

Cursul de Statistică socială şi analiza computerizata datelor face parte din categoria disciplinelor fundamentale de la Facultatea de Sociologie şi Asistență Socială din cadrul Universității Babeş-Bolyai din Cluj-Napoca. Cursul urmăreşte familiarizarea studentului cu conceptele si noțiunile necesare in analiza statistica a datelor culese in procesul de asistare socială şi cel al cercetării sociologice de teren. Se are in vedere dobândirea unor cunoştințe elementare de statistică descriptivă şi de testare a ipotezelor de cercetare. Prin exemplele şi aplicațiile utilizate se urmăreşte scoaterea în evidenta a rolului statisticii în profesia de asistent social. Tot aici se introduc şi primele noțiuni de utilizare a pachetului informatic SPSS *Statistical Package for Social Scientists*.

Organizarea temelor în cadrul cursului

Cursul este impartit in 11 module tratand fiecare subiecte specifice: introducere in statistica, distributiile de frecvente, graficele ataşate distribuțiilor de frecvente, indicatorii statistici ai tendintei centrale, indicatorii variabilitatii, distribuția normală, introducere in SPSS, noțiuni introductive de statistica inferentială, asocierea, compararea mediilor, şi corelația.

Nivelul de intelegere si, implicit, utilitatea informatiilor pe care le regăsiti in fiecare modul vor fi sensibil optimizate dacă, în timpul parcurgerii suportului de curs, veti consulta sursele bibliografice recomandate şi veți găsi răspunsurile corecte la întrebările pentru studiu de la sfârşitul fiecărui modul. Dealtfel, rezolvarea tuturor lucrărilor de verificare impune, cel puțin, parcurgerea referințelor obligatorii, menționate la finele fiecărui modul. În situația în care nu veți reuşi să accesați anumite materialele bibliografice, sau nu veți găsi răspunsuri la întrebări, sunteti invitați să contactați tutorii disciplinei.

Formatul si tipul activităților implicate de curs

Asa cum am mentionat mai sus prezentul suport de curs este structurat pe 11 module, necesare insuşirii graduale a tehnicilor de analiză a datelor. Parcurgerea acestora va presupune atât întâlniri față în

față (consultatii), cât şi muncă individuală. Consultatiile, pentru care prezenta este facultativa, reprezinta un sprijin direct acordat dumneavoastra din partea titularului si a tutorilor. Pe durata acestora vom recurge la prezentari ale informatiilor nucleare aferente fiecarui modul dar mai cu seama va vom oferi răspunsuri directe la întrebările pe care ni le veti adresa. În ceea ce priveşte activitatea individuala, aceasta o veti gestiona dumneavoastra si se va concretiza in parcurgera tuturor materilelor bibliografice obligatorii, rezolvarea întrebărilor pentru studiu de la sfâşitul modulelor, şi efectuarea exerciţiilor pentru acasă din cadrul lucrarilor de verificare. Reperele de timp si implicit perioadele in care veti rezolva fiecare activitate (exercitii, lucrari de verificare, etc) sunt monitorizate de catre noi prin intermediul calendarului disciplinei. Modalitatea de notare si, respectiv, ponderea acestor activitati obligatorii, în nota finala va sunt precizate în secţiunea politica de evaluare şi notare precum si in cadrul fiecărui modul.

Pe scurt, având în vedere particularităţile învăţământului la distanţă dar şi reglementările interne ale CFCID al UBB parcurgerea şi promovarea acestei discipline presupune antrenarea studenţilor în următoarele tipuri de activităţi:

a. consultaţii – pe parcursul semestrului vor fi organizate două întâlniri de consultaţii faţă în faţă; prezenţa la aceste întâlniri este facultativă;
b. laboratoare - se vor organiza pe parcursul semestrului intalniri faţă în faţă in laboratorul de informatica; prezenţa la aceaste întalniri sunt facultative;
c. două lucrări de evaluare (lucrari de control) care vor fi rezolvate şi, respectiv trimise tutorilor în conformitate cu precizările din calendarul disciplinei.
d. forumul de discuţii – acesta va fi monitorizat de echipa de tutori şi supervizat de titularul disciplinei.

Materiale bibliografice obligatorii

In suportul de curs, la finele fiecărui modul sunt precizate atât referinţele bibliografice obligatorii, cât şi cele facultative. Sursele bibliografice au fost astfel stabilite încât să ofere posibilitatea adâncirii nivelului de înţelegere si, implicit, de facilitare a aplicării fiecărei proceduri de analiză. Volumul „ *Metode statistice aplicate în ştiinţele sociale*" (1999, 2006), este referinţa obligatorie principală pentru cursul de „*Statistică socială şi analiza datelor"*; caracteristica sa definitorie e introducerea noţiunilor si conceptelor cu care lucrează statistica socială şi trecerea in revistă a celor mai utilizate metode de analiză a datelor in ştiinţele sociale. Cea de a doua lucrare de referinţă „ *Statistics for Social Workers"* (1987) permite accesul celor mai puţin obişnuiţi cu formalismul şi rigoarea procedurilor statistice la noţiunile şi limbajul specific al statisticii aplicate în cercetarea în asistenţă socială. In plus, al treilea manual obligatoriu "*Introducere în SPSS*" (2004) este un ghid de utilizare a pachetului informatic SPSS. Alternativ se poate utiliza sistemul de asistenţă on-line al SPSS sau volumul în limba engleză "*SPSS for Windows. Base System User's Guide*" (1992). Celor trei volume obligatorii se adaugă o serie de referinţe facultative, utile îndeosebi atunci când aveţi simţiţi nevoia unor explicaţii şi exemple suplimentare. Lucrările menţionate la bibliografia obligatorie se regăsesc şi pot fi împrumutate de la Biblioteca Facultăţii de Sociologie şi asistenţă socială din cadrul Bibliotecii Centrale „Lucian Blaga".

Materiale si instrumente necesare pentru curs

Optimizarea secvenţelor de formare reclama accesul studentilor la urmatoarele resurse:
- calculator conectat la internet (pentru a putea accesa bazele de date si resursele electronice suplimentare dar şi pentru a putea participa la secventele de formare interactivă on line)
- software SPSS instalat pe calculator
- imprimantă (pentru tiparirea materialelor suport, a temelor redactate, etc)
- acces la resursele bibliografice (ex: abonament la Biblioteca Centrală „Lucian Blaga")
- acces la echipamente de fotocopiere

Calendar al cursului

Pe parcursul semestrului II, în care se studiaza disciplina de față, sunt programate 3 întâlniri față în față (consultații) cu toți studenții; ele sunt destinate solutionarii, nemediate, a oricaror nelamuriri de continut sau a celor privind sarcinile individuale. Pentru prima întâlnire se recomandă lectura atentă a primelor şase module. La cea de a doua intalnire se vor efectua lucrari practice SPSS asistate, si este necesara parcurgerea prealabila a modulului al 7-lea. La cea de a treia intalnire se discuta ultimele patru module si se realizeaza o secventa recapitulativa pentru pregatirea examenului final. De asemenea in cadrul celor trei întâlniri studenții au posibilitatea de solicita titularului si/sau tutorilor sprijin pentru rezolvarea exercitiilor pentru acasă, in cazul in care nu au reuşit singuri. Pentru a valorifica maximal timpul alocat celor trei întâlniri studenții sunt atenționați asupra necesității suplimentării lecturii din suportul de curs cu parcurgerea obligatorie a cel puțin a uneia dintre sursele bibliografice de referința. Datele celor trei întâlniri sunt precizate în calendarul sintetic al disciplinei, vezi anexa A. În acelaşi calendar se regăsesc şi termenele la care trebuie transmise / depuse lucrările de verificare care totalizează cele 10 exerciții pentru acasă de la sfârşitul modulelor. Prima lucrare de verificare constă in exercițiile pentru acasă 1-6 (in care analizele statistice sunt efectuată manual), iar lucrarea a 2-a constă din exercițiile 7-10 (in care analizele statistice sunt efectuate cu SPSS).

Politica de evaluare si notare

Evaluarea finala se va realiza pe bază unui examen scris, tip grila, desfasurat in sesiunea de la finele semestrului II. Nota finală se compune din: a) punctajul obtinut la acest examen în proporție de 50% si b) aprecierea lucrărilor de evaluare pe parcurs – 50% (cate 25% fiecare lucrare).

Disciplina cuprinde două lucrari de verificare care vor fi transmise tutorelui la termenele precizate în calendarul disciplinei. Exercitiile care alcatuiesc cele două lucrări se regăsesc la sfârşitul fiecărui modul. Instrucțiuni suplimentare privind modalitățile de elaborare, redactare, dar şi criteriile de notare ale lucrărilor, vă vor fi furnizate de către titularul de curs sau tutori în cadrul întâlnirilor față în față.

Pentru predarea lucrarilor de verificare se vor respecta cu strictețe cerintele formatorilor. Orice abatere de la acestea aduce după sine penalizari sau pierderea totală a punctajului corespunzator acelei lucrări.

Evaluarea acestor lucrări se va face imediat după preluare, iar afişarea pe site a notelor acordate se va realiza la cel mult 2 săptămâni de la data depunerii/primirii lucrării. Daca studentul considera ca activitatea sa a fost subapreciata de către evaluatori atunci poate solicita feedback suplimentar prin contactarea titularului sau a tutorilor prin email.

Elemente de deontologie academica

Se vor avea în vedere următoarele detalii de natură organizatorică:

- Orice material elaborat de catre studenti pe parcursul activităților va face dovada originalitatii. Studentii ale căror lucrări se dovedesc a fi plagiate nu vor fi acceptati la examinarea finala.
- Orice tentativă de fraudă sau fraudă depistată va fi sancționată prin acordrea notei minime sau, in anumite conditii, prin exmatriculare.
- Rezultatele finale vor fi puse la dispoziția studentilor prin afisaj electronic.
- Contestațiile pot fi adresate în maxim 24 de ore de la afişarea rezultatelor iar solutionarea lor nu va depasi 48 de ore de la momentul depunerii.

Studenți cu dizabilități

Titularul cursului si echipa de tutori îşi exprima disponibilitatea, în limita constrangerilor tehnice si de timp, de a adapta conținutul şi metodele de transmitere a informațiilor precum şi modalitățile de evaluare (examen oral, examen on line etc.) în funcție de tipul dizabilității cursantului. Altfel spus, avem în vedere, ca o prioritate, facilitarea accesului egal al tuturor cursanților la activitățile didactice si de evaluare.

Strategii de studiu recomandate:

Date fiind caracteristicile învățământului la distanță, se recomandă studenților o planificare foarte riguroasa a secvențelor de studiu individual, coroborată cu secvențe de dialog, mediate de rețeaua net, cu tutorii si respectiv titularul de disciplina. Lectura fiecărui modul, găsirea răspunsurilor la întrebările de la sfârşitul modulelor şi rezolvarea la timp a exerciţiilor componente a lucrărilor de verificare garantează nivele înalte de înțelegere a conținutului tematic şi totodată sporesc şansele promovării cu succes a acestei discipline.

Modul 1

Introducere în statistică

1.1. Obiectul statisticii

Înainte de a încerca delimitarea obiectului statisticii de cel al altor ştiinţe, vom face câteva precizări legate de istoricul ei. Vom vedea astfel de unde derivă termenul "statistică" şi ce a însemnat aceasta la începuturi. Rezumând chiar de la început constatarea la care se ajunge în urma unui astfel de periplu istoric putem spune că, aproximativ în aceeaşi epocă – a doua jumătate a secolului al XVII-lea -, apar două curente de gândire, unul în Anglia şi altul în Germania. Ele au un element esenţial comun, anume preocuparea de a construi un corp de cunoştinţe pozitive asupra societăţii, dar se deosebesc destul de mult între ele, mai ales prin mijloacele utilizate în atingerea scopului propus, fiecare revendicându-şi dreptul de a fi considerat izvorul primar al statisticii. O judecată dreaptă va convinge pe oricine că ştiinţa statisticii, aşa cum arată astăzi, datorează aproape totul şcolii engleze, însă contribuţia universitară germană nu poate fi neglijată chiar şi numai pentru motivul că respectivei şcoli îi datorăm numele acestei ştiinţe.

Preocupările care se constituie în izvorul englez al statisticii formează curentul ce ar putea fi intitulat şi "abordarea cantitativă engleză", mai bine cunoscut sub numele de "aritmetică politică", după cum a botezat William Petty (1623-1687) noua disciplină socială, născută în 1662, atunci când John Graunt, un burghez cu preocupări ştiinţifice în timpul liber, a publicat un articol despre ceea ce astăzi s-ar numi "fenomenele demografice" din oraşul Londra.

Iniţiatorul celei de a doua linii, cea germană, este considerat Herman Conring (1606-1682), un savant cu un spirit enciclopedic, ocupând, pe rând, trei catedre la Universitatea Brunswick din Helmstaedt: cea de filosofie, cea de medicină şi apoi cea de ştiinţe politice. Paternitatea termenului "statistică" este atribuit însă unui alt profesor german, Gottfried Achenwall (1719-1772), care a trăit şi publicat un secol mai târziu.

Diferenţa dintre abordarea germană şi cea engleză constă în aceea că prima îşi propune în termeni mult mai clari să se constituie ca o ştiinţă globală asupra statului, aplecându-se asupra trăsăturilor caracteristice ale acestuia. Această viziune globalizantă avea drept finalitate să pună la îndemâna principilor instrumente eficace de conducere. Poate şi datorită dimensiunilor reduse (erau la vremea aceea aproximativ 300 de unităţi statale în Imperiul German), poate datorită altor cauze, cum ar fi slaba calitate a lumii academice germane, fărâmiţată şi ea într-o mulţime de universităţi provinciale, gânditorii germani nu au crezut de cuviinţă sau n-au fost capabili să privilegieze caracterul cantitativ-numeric al cunoştinţelor despre stat. Ei doreau mai curând să descrie verbal, să pună un diagnostic pe baza percepţiei şi intuiţiei şi să dea sfaturi privitoare la ameliorarea stării anumitor caracteristici ale statului.

Spre sfârşitul secolului al XVIII-lea, cele două direcţii ajung să se cunoască reciproc foarte bine – lucrările savanţilor germani fiind traduse în engleză – şi chiar se declanşează o concurenţă puternică între ele. Va învinge, la începutul secolului al XIX-lea, aritmetica politică şi, culmea ironiei, cu această ocazie, ea se va lepăda de propria-i denumire şi o va prelua pe cea a adversarului răpus: *Staatskunde* = ştiinţa despre stat.

Statistica s-a constituit deci ca *disciplină preocupată de culegerea, verificarea, prelucrarea şi interpretarea informaţiilor numerice referitoare la fenomenele sociale.*

Întemeietorii statisticii moderne se situează însă mai aproape, cu începere din a doua jumătate a secolului al XIX-lea, iar principalii reprezentanţi nu mai aparţin domeniului ştiinţelor sociale. Se produce astfel o mutaţie esenţială în concepţia generală despre obiectul statisticii. Vom aminti doar de Francis Galton (1822-1911), biolog şi antropolog, întemeietor, în 1901, al revistei Biometrika şi al ştiinţei numită biometrie, de marele antropolog şi psiholog Karl Pearson (1857-1936) şi de cel considerat cel mai mare dintre toţi, Ronald Aylmer Fisher (1890-1962), biolog şi cercetător într-o staţiune experimentală agrobiologică.

După cum se vede, nici una din cele trei personalităţi amintite, la care se pot adăuga şi altele, nu s-a preocupat de fenomene sociale. Începe astfel nu numai o fază calitativ superioară în dezvoltarea statisticii,

ci şi perioada de clarificare a conţinutului şi naturii acestei ştiinţe. Construită pe informaţia de tip social, statistica a fost considerată, de la început şi până la finele secolului al XIX-lea, o ştiinţă socială; abia multiplele aplicaţii în alte domenii (biologia, genetica, fizica etc.), iniţiate în aceea perioadă, au arătat cu claritate că simbioza statistică-demografie a fost un accident istoric, explicabil poate prin circumstanţele producerii lui, dar nu prin esenţa celor două discipline. Este evident că demografia nu poate să se constituie ca ştiinţă fără instrumentarul statistic, însă statistica putea să-şi forjeze aparatul prin aplicaţii în alte domenii.

În accepţiunea actuală, *statistica se axează pe tratarea informaţiilor numerice obţinute la nivelul unor mulţimi de entităţi, informaţii prelevate de la fiecare entitate în parte, dar care conduc la rezultate cu referinţă la ansamblu, şi nu la entităţile componente luate individual.*

1.2. Rolul statisticii în cercetarea ştiinţifică

În esenţă, ştiinţa este o metodă, un mod specific de a afla răspunsuri la întrebările pe care ni le punem. Principalele ei caracteristici sunt: căutarea unor reguli generale (legităţi), colectarea unor dovezi obiective, operarea cu afirmaţii controlabile, atitudine sceptică faţă de cunoştinţele acumulate, atitudine deschisă faţă de orice informaţii noi, creativitate şi transparenţă.

Utilizarea statisticii în asistenţa socială este impusă de exigenţele metodei ştiinţifice, ca metodă de culegere şi prelucrare şi interpretare a datelor. Dincolo de procedurile care o compun, metoda statistică este un concept abstract care poate fi mai uşor înţeles dacă îl raportăm la ceea ce înseamnă „metoda neştiinţifică", la modul în care ne fundamentăm cunoştinţele şi convingerile în viaţa de fiecare zi. În acest sens, se pot distinge trei modalităţi „neştiinţifice" de fundamentare:

- **Tradiţionalismul sau argumentul autorităţii**. Ceea ce ştim, învăţăm din contextul social sau cultural, se bazează pe obişnuinţe sau superstiţii. Uneori nici nu suntem conştienţi care este sursa cunoştinţelor noastre. Pur şi simplu, ştim şi orice argument contrar este respins înainte de a fi verificat în vreun fel. Dacă la baza unei informaţii cu care operăm se află şi o figură autoritară (părinte, profesor, „cei care au văzut"), atunci atitudinea necritică este şi mai consistentă.
- **Raţionalismul**. Baza acestuia este deducţia, pornirea de la un principiu general pentru a se ajunge la anumite concluzii specifice. Problema este că, pentru a obţine presupuneri corecte, atât principiul general cât şi presupunerile care decurg din acesta trebuie să fie corecte. Ceea ce nu se întâmplă întotdeauna. În plus, acest tip de raţionament nu aduce un plus de cunoaştere deoarece se bazează pe adevărurile care au condus, în trecut, la constituirea principiului pe care s-a bazat raţionamentul deductiv. Spre deosebire de acesta, raţionamentul inductiv urmează drumul de la specific la general şi permite noi explorări ale unui subiect. Inducţia este baza metodei ştiinţifice.
- **Empirismul**. Această modalitate se bazează pe concluzii extrase din experienţa de zi cu zi, din observaţia directă a faptelor. Ştiinţa este în mod esenţial empirică. Asta nu presupune că oamenii de ştiinţă nu se bazează pe logică sau raţionament deductiv. Dar acolo unde există îndoieli sau incertitudini, oamenii de ştiinţă acordă credit observaţiei şi nu deducţiei. Dar simpla observaţie, neînsoţită de metode adecvate de sistematizare a datelor şi de o analiză consistentă a semnificaţiei lor, nu este cu nimic mai bună decât un raţionament deductiv necritic.

Modalităţile cunoaşterii comune, enunţate mai sus, nu sunt prin ele însele lipsite de valoare. Deşi nu pot constitui argumente pentru enunţarea de concluzii, ele pot sluji în calitate de furnizoare de probleme şi ipoteze de cercetare. În nici un caz, însă, nu pot înlocui demersul doveditor al procedurilor inductive. În acest proces, statistica nu face decât să pună la dispoziţie un set de proceduri de calcul şi de raţionamente decizionale cu privire la semnificaţia datelor de cercetare. Rolul statisticii este acela de a descrie, de a face predicţii şi de a conferi credibilitate datelor de observaţie. Statistica pune ipoteza cercetării în faţa testului negaţiei, prin raportarea la un model aleator de distribuţie a valorilor măsurabile. Rezultatele care trec de acest test al negaţiei (ipoteza de nul) sunt considerate, sub rezerva unei erori probabilistice acceptate, ca fiind semnificative.

1.3. Utilizarea statisticii în profesia de asistent social

Asistenţii sociali au nevoie de cunoştinţe statistice. Aceasta nevoie este urgentată de creşterea numărului articolelor publicate în marile jurnale de specialitate, referitoare la cercetările efectuate şi de

tendința, ce reiese din manualele de specialitate, spre o muncă socială orientată empiric. Literatura de specialitate cuprinde frecvent rezultatele cantitative ale studiilor de cercetare. Articolele publicate folosesc de regulă argumente statistice, iar cititorii au nevoie de cunoştinţe în acest domeniu pentru a le înțelege. *Asistenții sociali trebuie să fie capabili să consume şi să înțeleagă literatura de specialitate.*

Cunoaşterea statisticii este necesară pentru luarea deciziilor, de exemplu atunci când o nouă metodă de terapie ar trebui sau ar putea fi folosită de către asistenți sociali în practica lor individuală. *Ca specialişti responsabili, asistenții sociali trebuie să-şi evalueze regulat eficiența.* Ei trebuie să se bazeze pe rezultate concrete şi nu doar pe propria lor perspicacitate şi intuiție; evaluarea trebuie să aibă baze empirice.

Pentru *a-şi aduce propria contribuție la cunoştințele de bază ale profesiei*, rezultatele cercetărilor trebuie comunicate şi altora. Pentru a fi credibili, ei trebuie să demonstreze că datele prelucrate au fost obținute în acord cu nişte reguli specifice. Aceste reguli nu sunt greu de înțeles; ele sunt bazate pe metode logice şi ştiințifice, aplicabile în situațiile de asistență socială, fie pentru practică, fie pentru cercetare.

Când asistenții sociali iau decizii practice, de multe ori se bazează pe simțul comun şi pe experiența anterioară. În orice caz ei nu trebuie să uite că "simțul comun" le-a spus poate că beneficiarii actelor caritabile preferă să nu muncească, că femeile nu sunt bune supervizoare, şi că pedeapsa (şi nu recompensa) îi face pe studenți să învețe mai rapid. O mai mare conştiinciozitate în folosirea metodelor de sondare şi a procedurilor statistice permite înlăturarea acestor prejudecăți.

Înțelegerea şi utilizarea procedurilor statistice în mod adecvat, permite asistentului social să depăşească obiectivele bazate exclusiv pe practica empirică.

1.4. Dificultăți şi riscuri în însuşirea metodei statistice

Dacă este să fim drepți, trebuie să recunoaştem că, în ciuda utilității ei incontestabile, statistica nu a fost niciodată disciplina preferată a studenților în asistență socială. Confruntarea lor cu această materie se dovedeşte a fi, uneori, o surpriză neplăcută la aflarea planului de învățământ universitar. Trebuie să acceptăm adevărul că majoritatea celor care vor sa acorde ajutor semenilor lor aflați in dificultate nu manifestă o simpatie deosebită pentru numere, formule şi algoritmi de calcul. De aici şi până la „fobia de statistică" nu este o distanță prea mare. Chiar dacă această „fobie" poate fi reală în unele cazuri, nu întotdeauna performanțele insuficiente în învățarea statisticii se datorează „statisticofobiei". Un curs introductiv de statistică aplicată este pe deplin accesibil oricărui absolvent de liceu, chiar şi celor care nu au urmat o secție de tip „real". Cu toate acestea anumite dificultăți nu pot fi negate, cele mai importante fiind:

- abundența de concepte noi, cu semnificații uneori dificil de înțeles în mod intuitiv;
- prezența unor concepte cunoscute din limbajul cotidian dar care au semnificații diferite în domeniul statisticii;
- dificultatea înțelegerii raționamentelor de tip probabilistic.

În altă ordine de idei, „naivitatea statistică" reprezintă un risc cel puțin la fel de mare ca şi „statisticofobia". Ea se referă la utilizarea în necunoştință de cauză a unor prelucrări statistice, fără a şti dacă sunt legitime sau nu şi fără a le înțelege semnificația. Acest tip de atitudine a apărut şi este favorizat de utilizarea programelor de calculator, care permit proceduri statistice sofisticate, altădată greu accesibile, şi care întrețin iluzia că ar putea fi aplicate în afara unei profunde cunoaşteri a rostului şi semnificației lor.

În fine, un alt tip de risc este cel dat de utilizarea statisticii cu scopul de a epata. „Epatantul statistic" abuzează de statistică, utilizează cât mai multe proceduri, face risipă de reprezentări grafice, de multe ori redundante, sau caută cu orice preț proceduri „exotice", rar utilizate şi cunoscute de puțină lume, chiar dacă ar ajunge la aceleaşi concluzii prin utilizarea unor proceduri „clasice". Pe scurt, se poate vorbi de „epatare statistică" ori de câte ori prelucrările trec înaintea raționamentului statistic şi, mai ales, înaintea demersului de cercetare. Statistica în asistență socială este, întotdeauna, un mijloc şi nu un scop în sine. Ea este un simplu instrument în atingerea unor obiective, dar un instrument fără care respectivele obiective nu pot fi atinse.

1.5. Noţiuni şi concepte utilizate în statistică

Individ statistic, variabilă

Noţiunile statistice elementare sunt cele de *individ statistic* şi *populaţie statistică*. Fără pretenţia de a defini exact nişte noţiuni care, probabil datorită caracterului lor primar, nici nu pot primi o definiţie în sensul strict al termenului, vom spune că ***indivizii statistici*** *sunt nişte entităţi elementare purtătoare de însuşiri*. Dintre aceste însuşiri una este comună şi exprimă natura însăşi a entităţilor respective, fiind atributul cu ajutorul căruia aceste entităţi sunt desemnate ca atare (oameni, grupuri umane, ţări, familii, aruncări ale zarului, extrageri ale unui număr de bile dintr-o urnă, măsurători efectuate cu un anumit instrument ale unei însuşiri fizice etc.); celelalte însuşiri sunt ***variabile***, adică au forme de manifestare diferite, la nivelul entităţilor diferite, şi tocmai studierea lor reclamă instrumentele statistice. Aşa de pildă, indivizii umani sunt diferiţi după o mulţime de însuşiri fizice (înălţime, greutate corporală, circumferinţa craniului etc.), psihice (temperament, inteligenţă etc.) sau sociale (statut social, nivel de educaţie etc.); familiile se pot deosebi după mărimea lor, după statusul social al membrilor ce le compun, după vârsta celui mai bătrân (tânăr) membru ş.a.m.d. În principiu, caracteristicile luate în considerare definesc specificitatea câmpului de interes al unei ştiinţe date.

Populaţie statistică, volum, eşantion

Mulţimea indivizilor de aceeaşi natură, care constituie ţinta unei investigaţii de tip statistic, formează ***populaţia statistică***. După cum se observă, termenii de individ şi populaţie trădează faptul că, iniţial, în sfera statisticii au fost cuprinse doar persoane şi populaţii umane. Din exemplele menţionate se vede că noţiunile s-au extins şi se aplică oricăror entităţi, respectiv mulţimi de entităţi, care nu mai au nimic de a face cu omul decât faptul că devin obiect de cercetare pentru acesta. Mai mult, respectivele entităţi *nu sunt neapărat obiecte materiale*, ele putând îmbrăca, aşa cum se vede din exemplele de mai sus, şi haina acţiunilor umane (aruncarea zarului, măsurarea repetată a aceluiaşi obiect etc.) sau alte forme.

Numărul indivizilor ce formează mulţimea statistică se numeşte **volumul** populaţiei. În principiu, tehnicile statistice lucrează, de preferinţă, cu *populaţii mari*, tocmai pentru ca felul în care se manifestă o însuşire cercetată să capete o anumită relevanţă la nivelul populaţiei, în sensul că o eventuală modificare a numărului de indivizi să nu producă schimbări esenţiale ale structurii de ansamblu.

Cu toată această cerinţă, în calculele statistice, volumul populaţiei intervine rareori în sensul de condiţie restrictivă, deci ca o valoare minimă a acestuia. De cele mai multe ori, volumul, **n**, al populaţiei intră în calculul indicatorilor statistici fără condiţii prealabile de amplitudine, deci indiferent dacă acesta este 10 sau 10 milioane. Relevanţa, expresivitatea şi, câteodată, chiar înţelesul mărimii astfel calculate sunt afectate negativ de o dimensiune prea redusă a populaţiei. A efectua calcule de statistică avansată sau a elabora modele explicative cu un număr mare de variabile atunci când se lucrează cu o populaţie de 10 indivizi este o absurditate logică evidentă. Aceasta nu înseamnă că, în general, instrumentul statistic nu se poate aplica şi la populaţii de dimensiuni mai reduse. O valoare medie, una mediană sau chiar indicatorii de dispersie ne dau o informaţie validă şi pentru o populaţie de 5-10 indivizi. Însă trebuie să înţelegem foarte clar că, cu cât intrăm în analize statistice mai sofisticate, cu cât implicăm simultan în analize mai multe însuşiri ale indivizilor (mai multe variabile) cu atât volumul populaţiei trebuie să fie mai ridicat.

Studierea populaţiilor mari ridică probleme practice, de culegere şi prelucrare a informaţiei, deosebit de dificile. Una dintre cele mai de seamă cuceriri ale statisticii, care are la bază teoria matematică a probabilităţilor şi care vine să rezolve tocmai aceste dificultăţi, o constituie utilizarea *procedurilor de selecţie*. Informaţia este culeasă doar de la o submulţime de indivizi – numită **eşantion** -, parte a cărei dimensiune poate fi extrem de mică în comparaţie cu populaţia totală, însă concluziile sunt generalizabile la întreaga populaţie.

Statistica descriptivă şi statistica inferenţială

În statistică sunt folosite două categorii generale: *statistica descriptivă şi statistica inferenţială*. ***Statistica descriptivă*** este folosită pentru a rezuma caracteristicile unui eşantion sau a unei populaţii. După ce datele despre membrii unui eşantion sau a unei populaţii au fost colectate, măsurătorile originale, sau

rezultatele (date brute), sunt organizate şi rezumate folosind tehnici ca acelea descrise în paragrafele următoare. Statistica descriptivă ne permite să derivăm informaţii noi din măsurătorile efectuate, care să descrie succint un eşantion sau o populaţie.

Statistica descriptivă se bazează pe măsurători făcute fie pe eşantioane fie pe o întreagă populaţie. Rezultatele ei nu se extind în afara subiecţilor particulari cercetaţi. Grafice, tabele şi numere descriptive cum ar fi media şi procentele sunt mai uşor de înţeles şi interpretat decât o listă lungă de date care raportează rezultatele măsurătorilor pentru fiecare variabilă şi în fiecare caz în parte. Scopul principal al statisticii descriptive este să reducă întreaga colecţie de date la termeni simpli şi mult mai uşor de înţeles, fără a distorsiona sau a pierde prea mult din informaţiile colectate.

Statistica inferenţială constă în proceduri pentru determinarea generalizărilor ce pot fi făcute, referitoare la caracteristicile unei populaţii, pe baza datelor colectate dintr-un eşantion. Statistica inferenţială este relevantă atunci când avem acces la un eşantion luat întâmplător dintr-o populaţie şi atunci când nu avem la îndemână toate datele care pot exista teoretic în întreaga populaţie.

Pentru a compara cele două tipuri de statistici, putem spune că statistica descriptivă furnizează informaţii despre un eşantion sau o populaţie prin date culese în mod real, în timp ce statistica inferenţială constă în proceduri care ne indică în ce măsură putem generaliza la o populaţie, informaţiile obţinute la cazurile studiate în mod real şi ce fel de constatări putem face. Numai statistica inferenţială ne permite să tragem anumite concluzii, pornind de la un eşantion ales întâmplător, despre o populaţie mai largă pe care nu am măsurat-o dar care a furnizat eşantionul. Ambele tipuri de statistici sunt accesibile cercetătorului în munca de asistenţă socială.

1.6. Variabile statistice şi măsurarea lor

Date, variabile şi valori statistice

Multe din activităţile asistenţilor sociali implică culegerea şi organizarea datelor despre diferitele aspecte ale profesiei şi despre clienţii lor. Cercetătorii din toate disciplinele folosesc forme structurate de păstrare a datelor. În profesia de asistent social se utilizează o largă varietate de metode pentru culegerea datelor, incluzând aplicarea chestionarelor, analizele de conţinut, experimentele sau observarea directă a comportamentului în condiţii naturale. În plus, se folosesc adesea date înregistrate de poliţie, materiale provenite de la recensăminte, de la registrele agenţiilor şi ale spitalelor. Informaţiile obţinute prin aceste metode precum şi prin alte metode, se numesc ***date (brute)***.

Datele sunt puncte de plecare pentru concluziile ştiinţifice, de aceea trebuie să fim foarte exacţi atunci când colectăm informaţii. Nici o concluzie rezultată din cercetarea ştiinţifică nu poate fi mai bună decât calitatea datelor pe care se bazează.

Există întotdeauna o limită a gradului în care datele pot fi culese şi folosite în mod real. În studiile de cercetare datele se limitează numai la anumite variabile. Aşa cum se amintea mai sus o ***variabilă*** este o caracteristică a individului statistic, o însuşire care prezintă interes ştiinţific. Dar *nu orice însuşire poate fi măsurată numeric*: uneori ea nu poate să reflecte mai mult decât o formă, stare sau tip, chiar dacă variază printre cazurile unei populaţii. Putem deci *să clasăm* indivizii în funcţie de astfel de variabile. Clasificarea este şi ea un fel de "măsurare". Fără să fim prea filozofici, putem spune că toate variabilele cu care au de-a face asistenţii sociali (atât în cercetare, cât şi în practică), sunt potenţial măsurabile. De exemplu, putem să măsurăm variabila "gen", clasificând clienţii noştri în bărbaţi şi femei, putem să măsurăm variabila "vârstă", clasificându-i după vârsta lor la ultima lor zi de naştere, putem măsura variabila "numărul clienţilor activi" într-un serviciu de asistenţă socială obţinând-o printr-o simplă numărătoare. Variaţiile potenţiale ale unei variabile date se referă la ***valorile variabilei***. De exemplu "femeie", "bărbat" sunt valori ale variabilei "gen".

Ca şi concept matematic, a măsura înseamnă a atribui numere sau simboluri unui aspect al realităţii obiective sau subiective, în funcţie de anumite aspecte cantitative sau calitative care le caracterizează. Modul în care sunt atribuite numere sau simboluri variabilelor statistice, poarta numele de „scale de măsurare"

1.7. Scale de măsurare a variabilelor statistice

Însuşirile indivizilor statistici se exprima (evaluează) sub diverse forme, cunoscute în literatura de specialitate ca "nivele de măsurare", "scale de măsură" sau pur şi simplu "scale". Principalele denumiri aplicate de statisticieni nivelelor de măsurare reflectă o precizie crescătoare: (1) nominal, (2) ordinal, (3) interval, (4) rapoarte.

Nivelul nominal

Nivelul de măsurare cu cea mai mică precizie este cel nominal. Cuvântul *nominal* sugerează repartiția numelor la categorii (valori), care sunt distincte unele față de altele. Măsurarea nominală este în esență un sistem de clasificare care implică categorizarea variabilelor în subclase. Valori diferite reflectă doar un tip diferit, nimic mai mult. Deoarece nu sunt implicate diferențieri de cuantum, nu este posibilă o ordonare a valorilor. Variabile cum ar fi genul, rasa, apartenența politică sunt de obicei considerate ca variabile de nivel nominal. Alte exemple sunt starea civilă, diagnosticul, ocupația, tipul de tratament.

Cerințele măsurării de nivel nominal sunt minime. O variabilă nominală poate avea două sau mai multe categorii, iar categoriile trebuie să fie distincte, să se excludă reciproc, şi să fie exhaustive. Termenii *distinct* şi *reciproc exclusiv* desemnează faptul că fiecare caz în parte trebuie să se potrivească numai uneia dintre categorii; termenul de *exhaustiv* indică faptul că trebuie să existe o categorie potrivită, pentru fiecare caz. De exemplu, există doar două clase pentru variabila nominală starea de viață - în viața sau decedat. Aceste categorii sunt în mod clar exhaustive şi reciproc exclusive, astfel încât fiecare persoană intră *numai într-una* din aceste două categorii (exhaustive), dar numai o categorie i se poate atribui (exclusiv).

Pentru măsurarea nominală, numerele (sau alte simboluri, cum ar fi literele) pot fi folosite ca semne convenționale pentru distingerea unei categorii fața de alta. Să presupunem că am divizat variabila "tipul de tratament" în trei categorii: terapie individuală, terapie de grup, terapie familială. În acest caz putem atribui eticheta Terapie 1 - terapiei individuale, Terapie 2 - terapiei de grup, Terapie 3 - terapiei de familie. Numerele pe care noi le-am folosit sunt doar etichete şi servesc numai pentru clasificare. Ar fi lipsit de sens în acest caz să spunem că 1 este mai mult sau mai puțin terapie decât 2 sau 3, sau să facem orice altă afirmație, care să dea orice conotație cantitativă categoriilor. În descrierea datelor de nivel nominal tot ceea ce putem face este să stabilim numărul cazurilor care intră în fiecare categorie.

Nivelul ordinal

Măsurarea la nivel ordinar nu numai că separă diferitele categorii de variabile, dar de asemenea face posibilă o ordonare a acestora de la înalt la scăzut, sau de la mai mult la mai puțin. Aceasta înseamnă ca le aranjează în categorii care intră într-o relație de supraordonare sau subordonare unele față de altele. Exemple de variabile care pot fi considerate de nivel ordinal sunt clasa socială, prestigiul ocupațional, nivelul de pregătire (bacalaureat, master, doctorat), evaluarea schimbării (considerabilă, într-o oarecare măsură, puțin, deloc), evaluarea eficacității tratamentului (foarte eficace, eficace, ineficace, foarte ineficace), evaluarea satisfacției clienților referitor la tratament (foarte satisfăcuți, satisfăcuți, nesatisfăcuți, foarte nesatisfăcuți), clasificarea gravității unei probleme (foarte gravă, gravă, simplă, foarte simplă), notarea cu calificative a unui examen (foarte bine, bine, suficient, insuficient).

Etichetele valorilor folosite în clasificarea de nivel ordinal face posibilă nu numai identificarea diferențelor dintre subclasele de variabile, dar şi a poziției lor relative. Printr-o clasificare de nivel nominal putem doar să exprimăm faptul că o valoare este diferită de alta. În clasificarea de nivel ordinal putem spune nu doar că o valoare este diferită de cealaltă dar şi faptul că prezintă o cantitate mai mare sau mai mică a variabilei.

Este important să notăm faptul că etichetele valorilor de nivel ordinal ale datelor nu indică cantități absolute şi nu asigură intervale egale între valori. De asemenea nu putem şti aranjarea exactă cuprinsă în toate categoriile. De exemplu putem avea o scală a pozițiilor sociale care sugerează clasificarea claselor sociale în acord cu un set de categorii aranjate de la Clasa 1 (superioară) la Clasa 5 (inferioară). Deoarece clasele nu reprezintă în mod necesar intervale egale, nu putem spune că Clasa 1 este la exact două clase interval superioară Clasei 3, sau că acest interval este exact aceeaşi distanță ca şi cea care separă Clasa 4 de Clasa 2.

De aceea, nivelul ordinal este cu un pas înainte față de cel nominal prin faptul că adaugă o diferențiere cantitativă a categoriilor şi face posibilă o observare ordonată. Diferitele valori ale unei variabile de nivel ordinal, oricum, nu indică cantitățile lor absolute, sau distanțele exacte dintre ele.

Nivelul interval

Nivelul interval, clasifică şi ordonează datele în funcție de valorile variabilelor; în plus le aranjează la distanțe egale. Spre deosebire de scalele ordinale, scalele de nivel interval au unitate de măsură, cum ar fi anul, gradul de temperatură, şi aşa mai departe. De aceea etichetele valorilor indică exact la ce distanță se află o valoare față de cealaltă. Cu variabilele de nivel interval putem arăta că un obiect are "mai mult" sau "mai puțin" dintr-o proprietate dată decât un alt obiect; de asemenea putem specifica cu câte unități are mai mult sau mai puțin.

Având distanțe egale între unități, valoarea 1 va fi la aceeaşi distanță de 4, ca şi valoarea 6 de 9, şi aşa mai departe. Într-un test creat pentru evaluarea inteligenței, cuantificat la nivel interval, diferența dintre IQ 100 şi IQ 105 *ar trebui* să reflecte aceaşi diferență ca cea dintre: IQ 115 şi IQ 120. Doi indivizi cu rezultatele de 50 şi respectiv 60 *ar trebui* să aibă aceeaşi diferență între rezultatele lor la testele de inteligență şi învățare ca cea dintre alți doi indivizi care au realizat punctajul de 80 şi respectiv 90. Deseori variabilele ca anomia, moralul grupului şi anumite atitudini sociale necesită să fie măsurate într-o scală de nivel interval. Accesul lor la statutul de variabilă de nivel interval este bazat pe munca cercetătorilor care au lucrat mulți ani la dezvoltarea unor instrumente adecvate de măsurare. Aceste instrumente au trecut printr-o serie de îmbunătățiri destinate creşterii gradului lor de precizie.

Scalele de nivel interval nu pot include o valoare de zero absolut (adică nu putem identifica un punct în care nu există nici o cantitate din variabila studiată). Aceasta înseamnă că nu putem spune că 2 este de două ori mai mult decât 1 - ci doar că este ***cu*** *o unitate mai mare*. De vreme ce citirea unei temperaturi de 0 grade pe un termometru nu reprezintă absența căldurii, o temperatură de 60 de grade nu este de două ori mai mare decât o temperatură de 30 de grade. Zero grade C nu este nimic mai mult decât un punct arbitrar ales pentru a primi eticheta valorii 0. Termometrele Celsius pot genera numai informații de nivel interval (spre deosebire de termometrele Kelvin, care au un punct 0 absolut).

Măsurătorile de nivel interval indică la ce distanță se află valorile unei variabile, unele față de altele. Ele nu indică magnitudinea absolută a valorii variabilei asociate fiecărui obiect sau persoană. Acest lucru este posibil numai folosind cel mai precis tip de măsurare, nivelul rapoarte.

Nivelul rapoarte

Existența unui punct zero natural, absolut, nonarbitrar constituie singura diferență între măsurarea la nivel interval şi cea la nivel rapoarte. De aceea numerele scalei de nivel rapoarte indică cantitatea reală a proprietăților măsurate. Cu astfel de scală putem stabili nu numai faptul că un obiect are cu atât mai multe unități dintr-o proprietate decât un al doilea obiect, dar şi faptul că primul obiect are *de atâtea ori* mai mult sau mai puține unități. Exemple de scale tip rapoarte sunt rata naşterilor sau a divorțurilor; numărul de copii într-o familie, numărul de comportamente observate pe parcursul unei anumite perioade de timp.

Punctul zero absolut are o semnificație empirică. Toate operațiile aritmetice sunt posibile: adunare, scădere, înmulțire împărțire. Aceasta permite utilizarea validă şi interpretarea completă a raportului dintre două rezultate. De exemplu, o țară cu rata totală a fertilității de 4,8 copii pe cuplu are o rată de două ori mai mare față de o țară cu rata de 2,4 copii pe pereche. În mod similar o familie cu un venit de 1.500.000 lei are un venit dublu față de o familie cu venitul de 750.000 lei.

Cele mai multe din datele folosite în asistența socială, atât în practică cât şi în cercetare nu furnizează măsurări de nivel rapoarte. Un mod de a testa existența unui nivel rapoarte este să ne gândim la posibilitatea valorilor negative. Dacă valorile negative pot fi logic repartizate (de exemplu temperatura de -25 grade Celsius) atunci măsurarea variabilei nu poate fi considerată mai mult decât de nivel interval.

Variabile calitative şi cantitative

Variabilele măsurate pe o scală de nivel nominal şi cele măsurate pe o scală de nivel ordinal formează categoria **variabilelor calitative,** numite şi **caracteristici statistice**. Atributul calitativ se conferă

datorită faptului că distincția între clase este una calitativă. Termenul calitativ este aici luat în contrast cu cel cantitativ, care are semnificație de măsurare propriu-zisă.

În contrast cu variabilele calitative vom avea **variabilele cantitative,** numite uneori simplu **variabile**. Ele se definesc prin aceea că proprietatea în cauză se concretizează, la nivelul fiecărui individ din populație, printr-o valoare numerică, rezultat al operației de măsurare. Numărul astfel atribuit este un număr cardinal şi redă intensitatea, magnitudinea cu care se manifestă însuşirea respectivă în cazul individului concret respectiv.

Variabile continue şi variabile discrete

Se numeşte „continuă" o variabilă de tip numeric (cantitativ, de tip interval/raport) care are un număr teoretic infinit de niveluri ale valorilor măsurate. Acest tip de variabilă poate lua, în principiu, orice valoare, permițând utilizarea zecimalelor. Exemple: timpul de reacție, înălțimea, greutatea, etc.

Se numeşte „discretă" o variabilă care prezintă un număr finit al valorilor pe care le poate lua (numărul persoanelor dintr-o familie, numărul de țigarete fumate zilnic).

Precizia măsurării şi indicatorii statistici

Cunoaşterea nivelului de măsurare al datelor de care dispunem serveşte la alegerea indicatorului statistic sau al testului statistic potrivit. Întotdeauna vom dori să utilizăm un indicator statistic adecvat sau să aplicăm testul statistic potrivit nivelului sau nivelelor de măsurare de care dispunem. Uneori testele statistice potrivite pentru datele de nivel interval, de exemplu, necesită satisfacerea unor cerințe suplimentare ale valorilor variabilei şi a modul în care sunt distribuite. Dacă aceste cerințe nu pot fi asumate, se pot utiliza testele statistice adecvate pentru variabile măsurate la nivel ordinal sau chiar la nivel nominal.

Nu este posibil să mergem în cealaltă direcție în ierarhia măsurării, de la mai puțin precis la mai precis. Dacă o variabilă este măsurată numai la nivel nominal, nu putem să o tratăm ca variabilă de nivel ordinal, de vreme ce lipseşte o ordonare a categoriilor variabilei. De aceea nu este corect să se aplice teste statistice create pentru măsurarea de nivel ordinal, pentru variabile considerate a fi de nivel nominal. În mod similar nu este potrivit să se aplice teste statistice destinate măsurării de nivel interval, variabilelor considerate a fi numai de nivel nominal sau ordinal.

Diferitele nivele de măsurare se referă la felul în care sunt conceptualizate şi măsurate fenomenele sociale şi nu la caracteristicile inerente ale fenomenelor în sine. După felul în care este conceptualizată şi operaționalizată o variabilă într-un studiu de cercetare, cum ar fi de exemplu "locul de reşedință", ea poate fi măsurată la nivelul nominal - dacă se foloseşte la indicarea numelui locului geografic al unei reşedințe -, la nivel rapoarte - dacă ea indică distanța de la acea reşedință la un punct specific de pe glob -, sau la nivel ordinal – dacă ea exprimă categoria administrativa a localității (sat, comuna, oras, municipiu).

Să luăm în considerare numărul variabil al şedințelor de tratament pe care le urmează un client într-o anumită perioadă de timp, de exemplu o lună. Dacă suntem interesați în numărul sau cantitatea tratamentului urmat într-o şedință, atunci putem să pretindem o variabilă de nivel rapoarte, de vreme ce poate să ia valorile 0, 1, 2, 3, etc. În orice caz aceeaşi variabilă poate fi utilizată ca indicator al cantității de ajutor terapeutic de care clientul are nevoie. Conceptul de "nevoie de ajutor" şi "frecvența şedințelor" de tratament efectuate nu sunt chiar aceleaşi. Numai pentru că clientul A a asistat la patru şedințe de tratament într-o lună dată, în timp ce un alt client B a asistat numai la două, nu suntem îndreptățiți să spunem că A a avut nevoie de un ajutor de două ori mai mare decât B. Variabila nu poate fi considerată de nivel rapoarte. Nici măcar de nivel interval nu este deoarece diferența cantității de ajutor de care cei doi clienții au avut nevoie, reflectată în diferența dintre două şi patru vizite, nu este aceiaşi ca şi acela reflectat în diferența dintre şase şi opt vizite. Este totuşi probabil că diferența numărului şedințelor de tratament reflectă o anumită diferență cantitativă în nevoia pentru ajutor terapeutic. De aceea variabila poate fi considerată de nivel ordinal.

Variabile dependente şi independente

În cercetarea ca şi în practica de asistenţă socială deseori suntem interesaţi în căutarea unei relaţii între două sau mai multe variabile. De obicei există o variabilă, numită ***variabilă dependentă***, pe care dorim să o examinăm. Vrem să ştim cât mai mult posibil despre factorii variaţiei variabilei dependente, cu scopul de a putea influenţa variaţia ei. Deseori, diferitele valori ale variabilei dependente pot fi legate de efecte cauzate de alţi factori. Aceşti factori fac trimitere la ***variabilele independente***.

Satisfacţia în muncă printre angajaţi poate să *depindă*, între altele, de felul de supervizare pe care îl primesc (satisfacţia = variabilă *dependentă*). Variaţiile în notele studenţilor la un curs opţional poate depinde de motivaţia studenţilor, de timpul afectat studiului, de capacităţile lor intelectuale (notele studenţilor = variabilă *dependentă*). Dacă un student reuşeşte să intre sau nu într-o relaţie de prietenie, poate *depinde* de capacităţile sale intelectuale, de posibilitatea de a-şi găsi prieteni, precum şi de competiţia în care intră cu alţi studenţi (găsirea de prieteni = variabilă *dependentă*). În toate exemplele de mai sus, diferitele valori ale variabilei dependente sunt logic dependente de celelalte variabile (independente), sau cel puţin am vrea să ştim dacă acest lucru este aşa. În fiecare exemplu, ar fi ilogic, dacă nu absurd, să credem că sensul relaţiei ar putea fi şi invers.

Asistenţii sociali, ca profesionişti, sunt de obicei mai interesaţi în a explica diferenţele dintre variabile cum ar fi satisfacţia în muncă, notele, succesul într-o relaţie de prietenie. De aceea, acestea vor fi variabile dependente în munca lor de cercetare. Sigur că cercetătorii care se ocupă de alte domenii ar putea căuta motivele variaţiilor în tipurile de supervizare, a variaţiilor în motivaţia la învăţare sau metodele folosite pentru căutarea unei relaţii de prietenie. În aceste studii altele ar deveni variabile dependente, iar cercetătorii ar putea folosi factori ca satisfacţia în muncă, notele sau succesul într-o relaţie de prietenie ca variabile independente. De aceea o *variabilă este considerată dependentă sau independentă, pe bază logică sau după punctul principal de investigare într-un studiu de cercetare dat.* Identificarea variabilei dependente şi a celei independente este necesară atât pentru formularea ipotezei cât şi pentru interpretarea rezultatelor testului statistic.

Observaţie. O relaţie de dependenţă nu implică în mod automat o relaţie de tip cauză-efect. Un simplu test statistic, chiar dacă este confirmat, nu este niciodată o dovadă că variaţia variabilei independente cauzează variaţia variabilei dependente.

Întrebări pentru studiu

1. Explicaţi prin ce diferă variabila calitativă de una cantitativă.
2. Daţi un exemplu de variabilă care poate fi dependentă într-un anumit studiu de cercetare şi independentă în altul.
3. Ce criteriu adiţional trebuie pus unei variabile pentru a putea fi considerată de nivel ordinal, criteriu care nu este cerut de o măsurare de nivel nominal?
4. Ce se cere în plus la măsurarea pe o scală de nivel rapoarte, faţă de variabilele măsurate la nivel interval?
5. Explicaţi de ce şi cum o variabilă "elev problemă" poate fi operaţionalizată de nivel nominal într-un studiu şi de nivel interval într-un alt studiu.
6. Ce furnizează în plus cercetătorilor statistica inferenţială faţă de statistica descriptivă?

Exerciţiu pentru acasă 1.

In jurul anului 2000 speranţa de viaţă feminină la naştere (e_o) în ţările Europei înregistrau următoarele valori (vezi tabelul de mai jos). Identificaţi individul statistic, variabila măsurată şi scala ei de măsură.

Ţara	e_o (ani)	Ţara	e_o (ani)	Ţara	e_o (ani)
Albania	75,50	Germania	80,53	Olanda	80,69
Austria	80,89	Grecia	80,43	Polonia	77,28
Belgia	81,10	Islanda	81,79	Portugalia	78,89
Bulgaria	74,77	Irlanda	79,04	România	74,20
Croaţia	76,01	Italia	81,65	Federaţia Rusă	72,90
Cipru	80,40	Letonia	75,50	San Marino	82,60
Rep. Cehă	78,05	Lituania	76,90	Rep. Slovacă	76,70

Danemarca	79,04	Luxemburg	80,69	Slovenia	79,07
Elveția	82,50	Macedonia	74,53	Spania	82,05
Estonia	75,61	Malta	80,07	Suedia	81,90
Finlanda	80,83	Marea Britanie	79,74	Turcia	71,20
Franța	82,24	Moldova	71,57	Ucraina	74,00
Georgia	75,90	Norvegia	81,28	Ungaria	75,25

Modul 2

Distribuții de frecvențe

După ce datele au fost culese ele trebuie organizate şi rezumate. Înainte ca prezentarea sub formă de tabel a unor date să poată fi realizată informația trebuie aranjată astfel încât ea să fie cât mai uşor de înțeles. Această secțiune va descrie cele mai utile moduri de tabelare a datelor.

2.1. Ordonarea datelor brute

Deseori, primul pas al procesului de rezumare a datelor este punerea lor într-o anumită ordine. Să presupunem, spre exemplu, că administratorul unei agenții de asistență socială doreşte să afle dacă agenția serveşte persoanele în vârstă ale comunității. (Se defineşte mai întâi ce înseamnă o persoană in vârstă, de exemplu cea având peste 50 ani). Administratorul decide să înregistreze vârsta clienților noi care apelează la serviciile agenției timp de o lună. Examinând fişele de înregistrare a acestor persoane, obține următoarele vârste.

32, 27, 26, 21, 37, 31, 32, 69, 26, 31, 37, 49, 31, 27, 37, 26, 49, 21, 27, 31

Deoarece datele nu sunt prezentate sub o formă anumită, ele constituie un sir simplu de date brute. Pentru a da un înțeles acestor date, ele trebuie aşezate într-o altă ordine, începând fie de la valoarea cea mai ridicată spre cea mai joasă, fie invers, ca în lista de mai jos.

21, 21, 26, 26, 26, 27, 27, 27, 31, 31, 31, 31, 32, 32, 37, 37, 37, 49, 49, 69

Acum se vede mult mai clar că cei mai tineri clienți au 21 ani (valoarea minimă) şi sunt doi la număr. Că cel mai vârstnic client are 69 ani (valoarea maximă) si este doar unul.

În situații reale, listele de date brute, chiar ordonate fiind, nu ne sunt de mare ajutor, mai ales atunci când numărul lor este mare. Pentru a elimina acest neajuns se foloseşte ***tabelul de frecvențe***.

2.2. Distribuții (tabele) de frecvențe

Frecvența este o noțiune fundamentală a statisticii, derivată din cea de individ, populație şi variabilă. Ea răspunde la întrebări de genul *Câți indivizi sunt la fel ?* sau *Câți aparțin aceleiaşi clase ?* Căutarea frecvențelor unei variabile înseamnă *aflarea numărului de indivizi statistici care populează fiecare clasă a variabilei*.

Prin urmare, dacă avem o caracteristică **X**, cu clasele $\mathbf{x_1, x_2, \ldots, x_s}$, şi dacă din cei **n** indivizi $\mathbf{k_1}$ sunt $\mathbf{x_1}$, $\mathbf{k_2}$ sunt $\mathbf{x_2}$, …., $\mathbf{k_s}$ sunt $\mathbf{x_s}$, atunci se realizează o corespondență biunivocă între mulțimea

$X = \{x_1, x_2, \ldots, x_s\}$ şi mulțimea $K = \{k_1, k_2, \ldots, k_s\}$

adică între mulțimea claselor şi cea a frecvențelor. Această ultimă mulțime, privită desigur prin prisma corespondenței cu **X**, poartă denumirea de **distribuție de frecvențe**. Distribuțiile de frecvențe se prezintă de regulă sub forma unor tabele numite **tabele de frecvențe**.

Distribuțiile de frecvențe, cum sunt tabelele 2.1 – 2.5, organizează în continuare datele exemplului început mai sus pentru a le face mai uşor şi mai rapid de înțeles. Vom prezenta cinci feluri de distribuții de frecvențe: tabelul de frecvențe absolute, tabelul de frecvențe cumulate, tabelul de frecvențe procentuale, tabelul de frecvențe procentuale cumulate, şi tabelul de frecvențe absolute şi procentuale. Le vom descrie pe fiecare pe rând.

Frecvențe absolute

Pentru a construi un tabel de frecvențe absolute, pur şi simplu numărăm de câte ori apare fiecare valoare pentru o anumită variabilă. Potrivind frecvențele pentru fiecare valoare, obținem o imagine de ansamblu a datelor. Aceasta ne permite să identificăm ordinea valorilor care apar, valoarea tipică, şi variația frecvențelor valorilor. O distribuție de frecvențe absolute ne permite să observăm la prima vedere cum sunt distribuite anumite valori ale unei variabile. O astfel de distribuție poate fi construită pentru date de orice nivel.

Partea stângă a Tabelui 2.1 arată faptul că vârsta clienților variază de la 21 la 69 sau că vârsta care apare cel mai des este 31 (de 4 ori). Coloana de frecvențe absolute din partea dreaptă a tabelului indică de câte ori apare fiecare valoare..

Tabelul 2.1. Tabelul distribuției de frecvențe absolute. Vârsta.

Vârsta (x_i)	Frecvențe (k_i)
21	2
26	3
27	3
31	4
32	2
37	3
49	2
69	1
Numărul total al clienților	20

Frecvențe cumulate

Dacă datele noastre sunt cel puțin de nivel ordinal (adică pot fi puse într-o ordine), atunci putem converti tabelul frecvențelor absolute (Tabel 2.1) într-un tabel de frecvențe cumulate. Un tabel al distribuției frecvențelor cumulate (notată cu F_i), ca şi Tabelul 2.2, este de fapt o extensie a unui tabel de frecvențe absolute. Tabelul 2.2 diferă de Tabelul 2.1 numai prin ultima coloană din dreapta care prezintă frecvențele cumulate ale vârstei clienților, de la cea mai mică (21), la cea mai mare.

$$F_i = k_1 + k_2 + \dots k_i \qquad i= 1, \dots, s$$

Aşa cum arată Tabelul 2.2, doi clienți aveau 21 ani, şi trei clienți aveau 26 ani. De aceea frecvența cumulată a vârstei de 26 ani sau mai puțin, este cinci (2+3). Un alt exemplu este faptul că 17 clienți (2+3+3+4+2+3) au vârsta de cel mult 37 ani.

Tabelul 2.2. Distribuție de frecvențe cumulate .Vârsta

Vârsta (x_i)	Frecvențe absolute (k_i)	Frecvențe cumulate (F_i)
21	2	2
26	3	5
27	3	8
31	4	12
32	2	14
37	3	17
49	2	19
69	1	20
Total	20	-

Frecvențe procentuale

Tabelul 2.3 este un tabel al distribuției de frecvențe procentuale care utilizează datele din Tabelul 2.2. Forma sa este identică cu cea a tabelului distribuției de frecvențe absolute, cum ar fi Tabelul 2.1, cu singura diferență că în loc de frecvențele absolute apar frecvențe procentuale (sau, mai simplu, procente). Frecvențele procentuale (p_i) se obțin prin raportarea frecvențelor absolute (k_i) la volumul populației (n) şi înmulțind apoi rezultatul cu 100.

$$p_i = k_i / n * 100 \quad i = 1, 2, \ldots, s$$

Putem vedea, de exemplu, din Tabelul 2.3, ca cele două persoane de 21 de ani, reprezintă 10% din numărul total de clienți, sau că procentul celor de 26 ani este de 15%, sau că o cincime (20%) dintre clienți au 31 ani. Totalul frecvențelor procentuale trebuie să fie întotdeauna 100% .

Tabelul 2.3 Tabelul de frecvențe procentuale. Vârsta.

Vârsta (x_i)	Procente (p_i)
21	10
26	15
27	15
31	20
32	10
37	15
49	10
69	5
TOTAL	100

Procente cumulate

Un tabel al distribuției de procente cumulate, cum este Tabelul 2.4, se construieşte asemănător tabelului de frecvențe absolute cumulate. Dacă un tabel al distribuției frecvențelor cumulate (Tabelul 2.2) prezintă numărul de indivizi cu vârste sub anumite limite, un tabel al distribuției de procente cumulate, prezintă procentul celor sub diferite vârste. Aşa cum apare în Tabelul 2.4, de exemplu, doar doi clienți au exact 32 de ani, împreună ei reprezintă 10% din totalul clienților, şi 70% din clienți au cel mult 32 ani.

Tabelul 2.4. Distribuție de procente cumulate. Vârsta

Vârstă	Procente	Procente cumulate
21	10	10
26	15	25
27	15	40
31	20	60
32	10	70
37	15	85
49	10	95
69	5	100
Total	100	-

Tabele de frecvențe absolute şi procentuale

Frecvențele absolute şi cele procentuale apar deseori într-un acelaşi tabel. Un tabel de frecvențe şi de procente, cum este Tabelul 2.5, este o combinație a două tabele: un tabel al distribuției frecvențelor cumulate (Tabelul 2.2) şi un tabel al distribuției procentelor cumulate (Tabelul 2.4).

Tabelul 2.5. Distribuţie de frecvenţe şi de procente. Vârsta

Vârsta	Frecvenţe	Frecvenţe Cumulate	Procente	Procente Cumulate
21	2	2	10	10
26	3	5	15	25
27	3	8	15	40
31	4	12	20	60
32	2	14	10	70
37	3	17	15	85
49	2	19	10	95
69	1	20	5	100
Total	20	-	100	-

2.3. Tabele de frecvenţe pentru variabile cu intervale de grupare

Uneori interpretarea tabelelor distribuţiei de frecvenţe poate fi dificilă din cauza aranjării inegale a valorilor ordonate ale mai multor variabile. În exemplul nostru variabila vârstă este distribuită în aşa fel încât apar pauze care duc la confuzii (21 la 26; 27 la 31; 32 la 37; 37 la 49; 49 la 69). Concluziile ar putea fi mai evidente dacă aceste date ar fi împărţite pe intervale de vârstă (de exemplu, 20-29; 30-39, etc.) şi rezumate într-o distribuţie de frecvenţe ca în tabelele 2.6 – 2.12. Valorile fiind grupate în intervale, şi frecvenţele corespunzătoare trebuiesc însumate.

Distribuţia de frecvenţe a variabilelor cu valori grupate pe intervale se utilizează nu numai atunci când operăm cu variabile discrete cu multe valori, dar şi atunci când avem variabile continue. Acesta este cazul des întâlnit la variabilele măsurate pe o scală de nivel interval sau rapoarte. Venitul anual al studenţilor, de exemplu, ar cuprinde o foarte lungă listă, mai ales dacă venitul ar fi măsurat, pentru fiecare individ, în dolari şi cenţi. Transformând observaţiile în categorii distincte, va fi mai uşor pentru noi să vizualizăm şi să rezumăm datele. De exemplu, venitul poate fi grupat în patru categorii: sub 600$; 601$ - 900$; 901$ - 1200$; 1201$ şi peste.

Când stabilim cum împărţim valorile in intervale trebuie să ţinem cont de câteva cerinţe: intervale să fie pe cât posibil egale ca lungime, să fie exhaustive (adică să acopere toate valorile posibile), să fie exclusive (adică orice valoare să se încadreze într-un singur interval fără nici un fel de dubiu), să nu fie prea multe (îngreunează interpretarea), să nu fie prea puţine (se pierde prea mult din precizie).

Ca şi în cazul distribuţiilor de frecvenţe cu valori discrete vom exemplifica cu cinci variante de tabele pentru distribuţii de frecvenţe pentru variabile cu intervale de grupare: tabel de frecvenţe (absolute); tabel de frecvenţe cumulate, tabel de procente, tabel de procente cumulate şi tabel de frecvenţe şi de procente.

Tabelul 2.6. Distribuţie de frecvenţe absolute. Grupa de vârstă

Grupa de vârsta	Frecvenţe
20-29	8
30-39	9
40-49	2
50-59	0
60-69	1
Total	20

Lungimea intervalelor în distribuţiile de frecvenţe (care este 10 în Tabelul 2.6) depinde atât de judecata noastră cu privire la ceea ce are importanţă, cât şi de numărul de observaţii ce trebuie clasificate. De obicei cu cât este mai mare numărul observaţiilor, cu atât este mai mare numărul intervalelor care pot fi folosite. Dacă se folosesc prea multe intervale (să zicem mai multe de 15) ele sunt atât de înguste încât informaţia este greu de acceptat, şi un rezultat important poate rămâne ascuns. Pe de altă parte, dacă se folosesc prea puţine intervale, foarte multă informaţie se poate pierde prin comasarea unor observaţii diferite. Intervalele nu trebuie să fie atât de largi încât două valori incluse să aibă între ele o diferenţă

considerată importantă. De exemplu dacă nu este considerată ca importantă o diferență de 200$ în venitul anual al unei familii, dar este considerată importantă o diferență de 300$, putem alege următoarele intervale: sub 200$, 201$ - 400$; 401$ - 600$ etc.

Tabelul 2.7. Distribuție de frecvențe cumulate. Grupa de vârstă

Vârsta	Frecvențe	Frecvențe cumulate
20 - 29	8	8
30 - 39	9	17
40 - 49	2	19
50 - 59	0	19
60 - 69	1	20
Numărul total de clienți	20	

Tabelele de procente prezintă atât frecvențele absolute ale fiecărui interval cât şi procentele reprezentate de indivizii din fiecare interval în raport cu totalitatea cazurilor. De aceea, aşa cum indică Tabelul 2.8, opt clienți aveau vârsta de la 20 la 29 ani; aceşti opt clienți reprezintă 40% din întregul eşantion. Aşa cum vom arăta puțin mai jos, o asemenea prezentare este utilă mai ales atunci când comparăm valorile unei variabile din două sau mai multe grupuri de mărimi diferite.

În Tabelul 2.8 putem vedea procentul clienților din fiecare grupă de vârstă. 45% , adică 9 dintre clienți, sunt incluşi în categoria de vârstă 30 - 39 ani. Prin adunare obținem că 85% (40% + 45%) sunt în categoria 20 - 39 de ani.

Tabelul 2.8. Distribuție de procente. Grupa de vârstă.

Vârsta	Frecvențe	Procente
20 - 29	8	40
30 - 39	9	45
40 - 49	2	10
50 - 59	0	0
60 - 69	1	5
Total	20	100

Datele prezentate în Tabelul 2.9, al procentelor cumulate, indică faptul că îngrijorarea de la început a administratorului agenției era justificată. Adică, majoritatea (95%) clienților care solicită serviciile agenției au vârsta sub 50 ani. Această concluzie este mult mai precisă şi mai convingătoare, decât afirmația că "mai mulți" dintre clienți, se încadrează între limitele vârstei de 20 - 39 ani, decât între 40 şi 69 ani.

Tabelul 2.9. Distribuție de procente cumulate. Grupa de vârstă

Vârsta	Procente	Procente cumulate
20 - 29	40	40
30 - 39	45	85
40 - 49	10	95
50 - 59	0	95
60 - 69	5	100
Total	100	-

Tabelul 2.10. Distribuție de frecvențe absolute şi procentuale. Grupa de vârstă.

Vârsta	Frecvențe	Frecvențe Cumulate	Procente	Procente Cumulate
20 - 29	8	8	40	40
30 - 39	9	17	45	85
40 - 49	2	19	10	95
50 - 59	0	19	0	95
60 - 69	1	20	5	100
Total	20	-	100	-

2.4. Utilizarea frecvențelor absolute şi relative în comparații

Distribuțiile de frecvențe cumulate sunt utile atunci când suntem interesați de poziția cumulativă a unei valori. Să presupunem, spre exemplu, că administratorul unei mari organizații de servicii sociale, dorește să studieze problema absenteismului nemotivat al personalului. El ar dori să identifice specificul fiecărui anotimp pentru a putea îmbunătăți situația prin tactici noi pentru vacanță şi pentru concediul anual. Ar putea fi utile în acest caz fie un tabel de frecvențe cumulate, cum este Tabelul 2.11, sau un tabel de procente cumulate, cum este Tabelul 2.12.

Aceste două tabele indică faptul că în luna aprilie au apărut 30 de absențe (Tabelul 2.11), sau numai 15% (Tabelul 2.12) din totalul absențelor pe o perioadă de patru luni. Cantitatea cumulată a absențelor a fost de 200 de zile (Tabelul 2.11).

Tabelul 2.11. Tabelul distribuției de frecvențe cumulate
Absențele nemotivate pe fiecare lună la agenția XYZ

Luna	Frecvențe	Frecvențe cumulate
Aprilie	30	30
Mai	40	70
Iunie	60	130
Iulie	70	200
Total	200	-

Tabelul 2.12. Tabelul distribuției de procente cumulate
Absențele nemotivate pe fiecare lună la agenția XYZ (din Tabelul 2.11)

Lună	Procente	Procente cumulate
Aprilie	15	15
Mai	20	35
Iunie	30	65
Iulie	35	100
Total	100	-

Numai 35% (Tabelul 2.12) din absențe sau petrecut în primăvara târzie, în timp ce celelalte 65% (30% + 35%) s-au petrecut în timpul lunilor iunie şi iulie. De aceea din cele două tabele administratorul agenției a putut scoate în evidență sezonalitatea absenteismului.

Frecvențele cumulatepentru variabile cu intervale de grupare (mai ales cele procentuale) sunt utile în special atunci când dorim să comparăm distribuțiile a două sau mai multe seturi de date, în termeni de câți (ce procente) de indivizi se situează sub o anumită valoare. Pot fi folosite, spre exemplu, pentru a obține o primă evaluare a posibilului succes al unui îndrumător de studiu. Am putea să privim rezultatele persoanelor care au utilizat îndrumătorul (grupul experimental) şi comparativ cu rezultatele persoanelor care nu l-au folosit (grupul de control). Tabelele de frecvențe cumulate care arată diferențele dintre rezultatele celor două grupuri sunt Tabelele 2.13 şi 2.14.

Tabelul 2.13. Tabelul distribuției de frecvențe şi procente cumulate.
Rezultatele obținute la examen de grupul celor care au folosit indrumătorul (n = 300)

Puncte	Frecvențe	Frecvențe Cumulate	Procente	Procente Cumulate
50 - 59	0	0	0	0
60 - 69	30	30	10	10
70 - 79	120	150	40	50
80 - 89	90	240	30	80
90 - 100	60	300	20	100

Tabelul 2.14 Tabelul de frecvențe cumulate şi de procente cumulate grupate:
Rezultatele obținute la examen de grupul de control (n = 200)

Puncte	Frecvențe	Frecvențe Cumulate	Procente	Procente Cumulate
50 - 59	0	0	0	0
60 - 69	30	40	15	20
70 - 79	80	120	40	60
80 - 89	70	190	35	95
90 - 100	10	200	5	100

Putem vedea în Tabelul 2.14, că 20% dintre asistenții sociali (40 la număr) din grupul de control au obținut la examen un rezultat de 69 puncte sau mai puțin. Tabelul 2.13 indică faptul că doar 10% dintre asistenții sociali (30 în valoare absolută) care au utilizat îndrumătorul pentru studiu au obținut un rezultat asemănător. A trebuit să ținem cont în comparare că cele două grupuri erau alcătuite dintr-un număr inegal de subiecți (n), 300 din grupul experimental şi 200 din grupul de control. Numai folosind procentele am putut compara două grupuri de mărimi diferite.

Din perspectiva cercetării, cele două grupuri de 200, şi respectiv de 300, subiecți pot fi comparate în mod rezonabil. Compararea procentelor simplifică interpretarea datelor dar trebuie să fim prudenți cu privire la compararea procentelor între două grupuri a căror mărime diferă mult. Un astfel de procedeu în realitate poate mai degrabă să distorsioneze decât să clarifice datele pentru cititor. Exemplul de mai jos va arăta cum se poate întâmpla acest lucru.

Să presupunem că un asistent social se ocupă de 10 clienți în timp ce un altul de 100 de clienți. Cel care are 10 clienți termină două cazuri pe parcursul unei luni, iar celălalt termină şase cazuri. Primul asistent social raportează un procent de 20% (2/10 =20%) a cazurilor încheiate, în timp ce celălalt raportează numai 6% (6/100 = 6%) cazuri încheiate. Privind numai datele procentuale (20% contra 6%), apare o diferență semnificativă între ratele relațiilor încheiate la cei doi asistenți sociali (20 este mai mare decât 6 de mai mult de trei ori). Este această diferență atât de semnificativă?

Pentru a evita o falsă impresie numerele reale pe care se bazează procentele, trebuie întotdeauna raportate împreună cu procentele dacă numărul de cazuri din ambele grupuri este mare, iar grupurile sunt inegale în mărime. În general procentele sunt lipsite de semnificație, dacă nu total derutante, atunci când sunt raportate datele unor eşantioane mici. Ar fi cel mai bine ca ele să nu fie raportate, de vreme ce numerele mici sunt destul de uşor de înțeles.

Întorcându-ne la exemplul cu examenul, putem să observăm că procentele cumulate fac posibilă calcularea, cel puțin aproximativă, **centilei** pentru fiecare individ. Centila unei valori indică procentajul cazurilor a căror valori depăşesc sau se situează sub nivelul valorii date. Să presupunem că un anumit individ din grupul experimental a obținut în cadrul examinării un rezultat de 90 puncte, după ce a folosit îndrumătorul de studiu. Din Tabelul 2.13 putem vedea că persoanele care au obținut un rezultat mai mic decât el reprezintă 80% din toate persoanele grupului experimental - adică individul se situează la nivelul centilei 80%. Frecvențele procentuale cumulate ne permit să privim un rezultat individual raportându-l la celelalte rezultate din grup.

Întrebări pentru studiu

1. Ce trebuie să avem in vedere atunci când determinăm lungimea intervalelor de grupare?
2. Ce informație în plus apare la distribuția de frecvențe cumulate şi care nu este prezentă la distribuția de frecvențe absolute?
3. Ce tip de distribuție de frecvențe ne va spune cel mai rapid, cât la sută din 396 studenți are vârstă de 25 ani sau mai puțin?
4. Ce tip de distribuție de frecvențe ne va spune câți clienți beneficiari de ajutor social dintr-o localitate au mai puțin de patru copii?
5. Unde este greşeala la o distribuție de frecvențe grupate după vârstă, care foloseşte categoriile de vârstă între 20 şi 30, 30 - 40, 40 - 50, 50 - 60, 60 - 70 şi peste 70 ani?
6. Într-un studiu care încearcă să relateze despre modul de obținere a unei slujbe, de ce este contraindicat să se grupeze variabila "număr de interviuri" de la 1 - 10, 11 - 20, şi peste 20?
7. Ce ne spune un rezultat care cade pe al 73-lea procent cumulat, despre performanțele unui student comparație cu ceilalți colegi ai lui?

Exerciţiu pentru acasă 2.

Construiți un tabel de frecvențe pentru datele de la Exercitiul pentru acasa 1
Adăugați inclusiv frecvențele procentuale şi frecvențele procentuale cumulate.

Modul 3

Prezentarea grafică a distribuțiilor de frecvențe

Este greu de sesizat toate semnificațiile unui tabel, dar acestea pot fi comunicate aproape imediat prin imagini. Atunci când datele se prezintă în mod eficient sub formă grafică, ele sunt mai degrabă văzute decât studiate; ele sunt sesizate ca un întreg şi nu sub formă de fragmente aşa cum se poate întâmpla chiar şi atunci când tabelele statistice sunt examinate cu atenție. Din acest motiv prezentarea grafică este utilă, în mod special pentru a afişa rezultatele cantitative ale unui studiu de cercetare care implică discuții statistice. Este adevărat faptul că reprezentarea grafică poate fi modificată uşor astfel încât să ducă la concluzii eronate. Dar acest avertisment nu trebuie să ne descurajeze privind folosirea lor.

O altă piedică în folosirea reprezentărilor grafice este faptul că în general ele renunță la detalii şi uneori la acuratețe în efortul de îmbunătățire a comunicării. Acest sacrificiu este justificat numai atunci când (1) concluzia este foarte importantă, (2) subiectul poate fi foarte eficient reprezentat în imagini, (3) auditoriul nu este orientat către cercetare şi astfel nu ar putea înțelege un tabel la fel de uşor ca o reprezentare grafică sau (4) este esențial să se treacă la subiect rapid pentru că s-ar putea întâmpla ca auditoriul să îşi piardă răbdarea şi interesul.

Graficele, diagramele şi figurile se folosesc însă mai frecvent decât ar putea rezulta din cele spuse anterior. Ele servesc nu numai comunicării informațiilor către alte persoane, dar şi cercetătorului în munca lui de explorare, de căutare sau de analiză a datelor colectate. Mai ales dacă sunt obținute cu ajutorul calculatorului şi vor putea fi multe la număr cercetătorul le va folosi ca instrumente de cercetare la propriu.

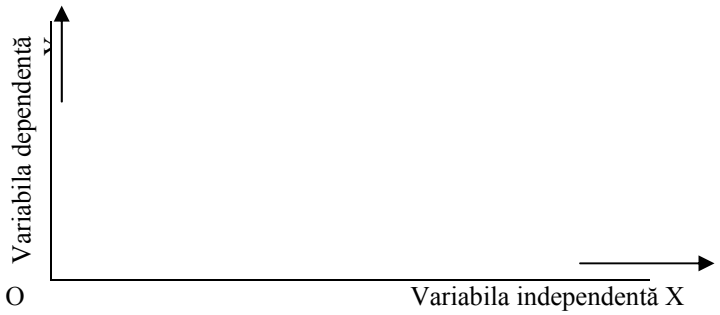
Graficele prezintă variabilele, conceptualizate şi operaționalizate într-o fază anterioară a cercetării, într-o formă uşor de priceput.

Figura 3.1. Sistem de axe pentru grafice ataşate distribuțiilor de frecvențe

Distribuțiile de frecvențe, ca de altfel toate tabelele statistice, pot fi transpuse într-o multitudine de forme grafice al căror rol este acela de a face mai perceptibile trăsăturile distribuției. Cele mai des utilizate grafice sunt bazate pe sistemul perpendicular de axe de coordonate. Pe axa orizontală se vor figura clasele sau valorile variabilei de reprezentat, iar pe axa verticală frecvențele - absolute, relative sau cumulate – ca în figura 3.1.

Reprezentările grafice ale datelor pentru două variabile sunt bazate pe două linii perpendiculare, aşa cum apare în figura 3.2. Linia verticală este cunoscută sub numele de axa Y, iar cea orizontală, axa X. În graficele care prezintă relația dintre două variabile, iar aceste pot fi identificate ca dependente sau independente, axa Y este folosită pentru a reprezenta valorile variabilei dependente, iar axa X pentru a reprezenta valorile variabilei independente. Valorile pentru fiecare variabilă (dependentă şi independentă) sunt înşirate pe o grilă uniformă, de-a lungul axei corespunzătoare. Punctul unde axa X şi Y se întâlnesc se numeşte "punct de origine". Începând cu punctul de origine, valorile variabilelor dependente şi independente sunt plasate de-a lungul axelor respective, într-o ordine ierarhică, la intervale egale. Valorile pentru variabile dependente şi independente dintr-un anumit grafic pot sau nu să aibă intervale de lungimi identice. Valorile pentru variabila dependentă pot fi marcate la un interval de cinci unități, spre exemplu, în timp ce valorile variabilei independente pot fi marcate la un interval de zece unități.

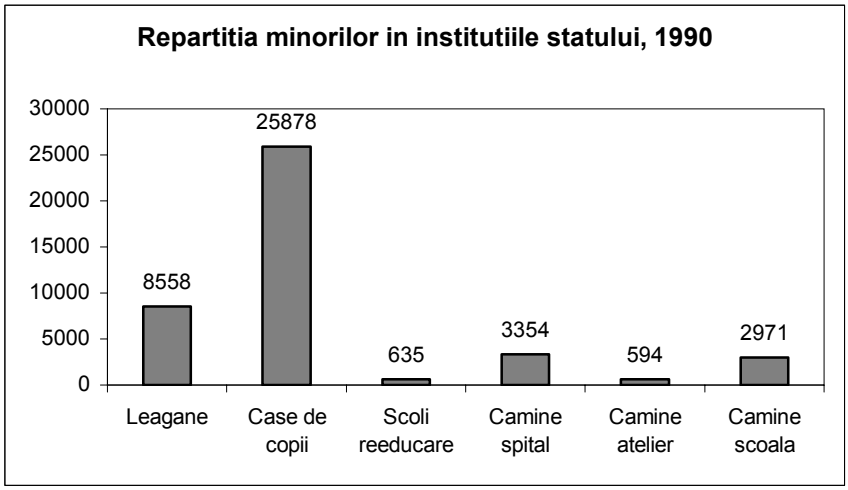
Figura 3.2. Sistem de axe destinate evidenţierii relaţiilor dintre două variabile



Deoarece există multe tipuri de grafice, vom discuta numai cinci dintre ele, utilizate frecvent în cercetarea de asistenţă socială: diagrama de bare, histograma, diagramă circulară, poligonul de frecvenţe, diagrama de împrăştiere. Primele două vor fi prezentate împreună deoarece se utilizează într-o manieră similară.

3.1. Diagrama de bare şi histograma

Figura 3.3.Exemplu de diagramă de bare



Una din cele mai utilizate forme de reprezentare a distribuţiilor de frecvenţe este diagrama de bare. Pe axa orizontală se construiesc la distanţe egale bare de lăţimi egale şi de înălţimi proporţionale cu frecvenţele. Barele sunt desenate astfel încât ele să nu se atingă (vezi figura 3.3) şi pentru a sugera diferenţele calitative şi nu cantitative ale valorilor variabilei. Acest tip de grafic este ideal pentru distribuţiile de frecvenţe a variabilelor de nivel nominal şi ordinal.

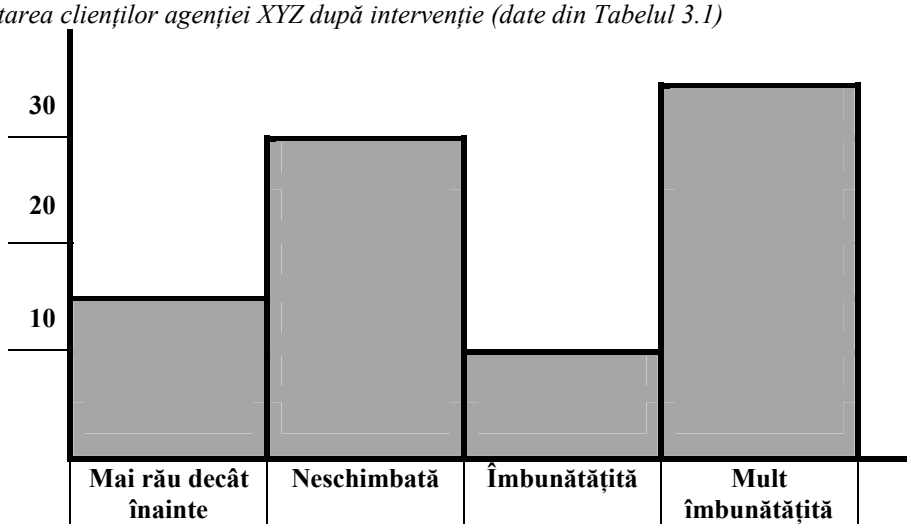
Datele măsurate pe o scală ordinală pot fi reprezentate eficient şi pe un alt tip de grafic, numit histogramă. O **histogramă** seamănă mult cu un grafic de bare dar aici atingerea barelor este permisă. Histogramele se mai folosesc atunci când se reprezintă datele pentru o variabilă cu intervale de grupare, pentru a pune în evidenţă diferenţe cantitative.

Figura 3.4 este un exemplu de histogramă, derivată din tabelul 3.1, şi afişează frecvenţele unei variabile de nivel ordinal - starea pacienţilor care au părăsit agenţia în care au primit ajutor.

Tabelul 3.1 Tabel de frecvenţe pentru starea clienţilor agenţiei XYZ, după intervenţie

Rezultat	Frecvenţe
Mai rău decât la admitere	15
Neschimbată	30
Îmbunătăţită	10
Mult îmbunătăţită	35
TOTAL	90

Figura 3.4. Exemplu de histogramă pentru o variabilă ordinală
Starea clienţilor agenţiei XYZ după intervenţie (date din Tabelul 3.1)



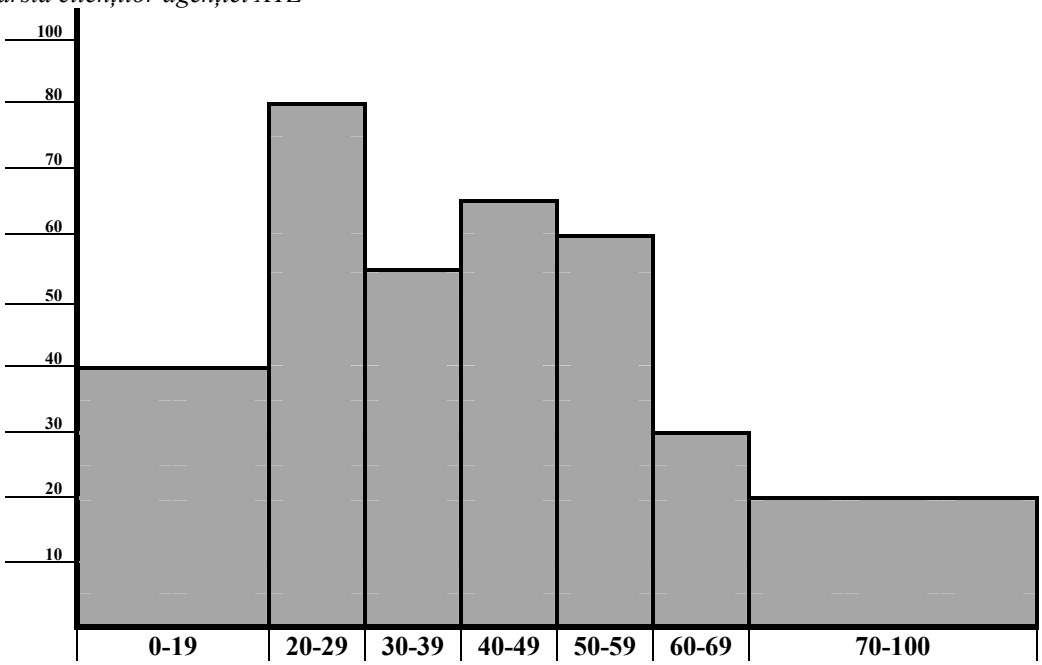
Pentru variabilele de nivel interval sau rapoarte, mai ales dacă distribuţiile de frecvenţe folosesc intervale de grupare, se utilizează histogramele şi nu diagramele de bare.

Un caz special, când histograma se deosebeşte de diagrama de bare nu doar prin simplul fapt că barele se ating, este atunci când intervalele de grupare a variabilei nu sunt egale ca lungime. Lăţimea barelor trebuie să fie proporţională cu lungimea intervalului. Dacă intervalele au lungimi egale, barele vor fi de aceeaşi lăţime. Dacă intervalele nu sunt egale atunci barele vor avea lăţimi diferite (ca în tabelul 3.2. şi figura 3.5).

Tabelul 3.2 Tabel de frecvenţe pentru vârsta clienţilor agenţiei XYZ

Intervalul de vârstă	Număr de clienţi
0- 19	80
20-29	80
30-39	55
40-49	65
50-59	60
60-69	30
70-100	60
TOTAL	430

Figura 3.5. Exemplu de histogramă pentru o variabilă cantitativă.
Vârsta clienților agenției XYZ



Înălțimea barelor, ca şi în cazul diagramei de bare, va fi proporțională cu frecvențele. Dacă intervalele de grupare nu sunt egale înălțimile barelor pentru intervalele mai lungi trebuie diminuate proporțional cu lățimile. Acest lucru este foarte important pentru că în cazul histogramelor contează de fapt suprafața barelor şi nu înălțimea lor ca în cazul diagramelor de bare.

Graficele de bare şi histogramele pot deveni un exercițiu de creativitate. Din nefericire, cu cât sunt mai creative, cu atât ele sunt mai dificil de interpretat. Barele orizontale în graficele cu bare pot fi extinse la dreapta sau la stânga, sau în ambele direcții simultan, dar astfel graficul devine repede mult mai greu de înțeles deoarece prezintă prea multe informații.

Uneori suntem tentați să folosim cât mai complex şi mai creativ diagramele de bare şi histograme, dar trebuie întotdeauna să ne întrebăm dacă graficul ajută în mod real cititorul să înțeleagă informațiile noastre? Sau vor avea un efect nedorit în confuzionarea cititorilor raportului de cercetare.

3.2. Diagrama circulară

Câteodată diferitele valori ale unei variabile sunt privite ca părți ale unui întreg şi categoriile pot fi mai bine ilustrate ca segmente ale întregului. Dacă dorim să arătăm cum se împarte bugetul familial al clientului pentru: hrană, locuință, haine, recreare, putem etala bugetul ca un întreg - ca un cerc. Diagramele circulare sunt cercuri sau discuri împărțite în sectoare de cerc reprezentând fracțiuni ale unui întreg. Deoarece diagramele circulare pot ilustra doar distribuția unei singure variabile, ele nu se pot folosi în sistemul de axe perpendiculare X-Y .

Figura 3.6.Exemplu de diagramă circulară

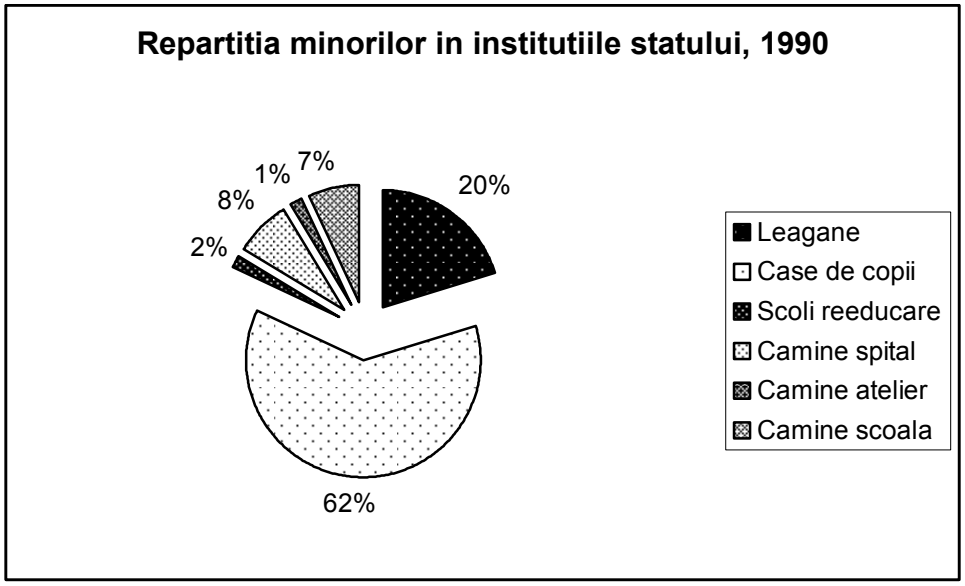


Figura 3.6 este un exemplu de diagramă circulară care arată repartiția minorilor instituționalizați în 1990 în România printre diferitele tipuri existente de instituții. Sectoarele de cerc sunt proporționale cu frecvențele întâlnite, fie că le considerăm pe cele absolute sau pe cele procentuale. Astfel, de exemplu, pentru minorii din casele de copii s-a alocat un sector de cerc de 73,4 grade adică 20,4% (25870 / 41982 *100 = 20.4) din totalul de 360 grade cât are un cerc întreg. Într-adevăr:

360 * 20,4 / 100 = 73,4

Diagramele circulare necesită aceleaşi date ca şi diagramele de bare. Orice informație care poate fi categorisită ca fiind de nivel nominal, poate fi reprezentată prin diagrame circulare. Principalul avantaj este că furnizează, rapid şi în mod vizual, aprecieri asupra informațiilor care pot face ca rezultatele constatărilor să fie mai pline de înțeles. Principalul dezavantaj al acestui tip de grafic este că ele nu pot folosi eficient multe valori diferite ale variabilei; adesea graficele devin prea complicate sau chiar neinteligibile în cazul în care sunt necesare mai mult de cinci sau şase diviziuni ale cercului.

3.3. Poligonul de frecvențe

Poligoanele de frecvență, sau **liniile poligonale** (cum sunt denumite în majoritatea produselor software de reprezentare grafică a datelor) sunt similare histogramelor şi sunt destinate descrierii formei distribuțiilor de frecvențe. Dacă am lua un creion şi am pune un punct în vârful şi la mijlocul fiecărei bare verticale dintr-o histogramă şi apoi am conecta punctele între ele printr-o linie dreaptă, vom realiza un poligon de frecvență.

Figura 3.7. Exemplu de poligon de frecvențe

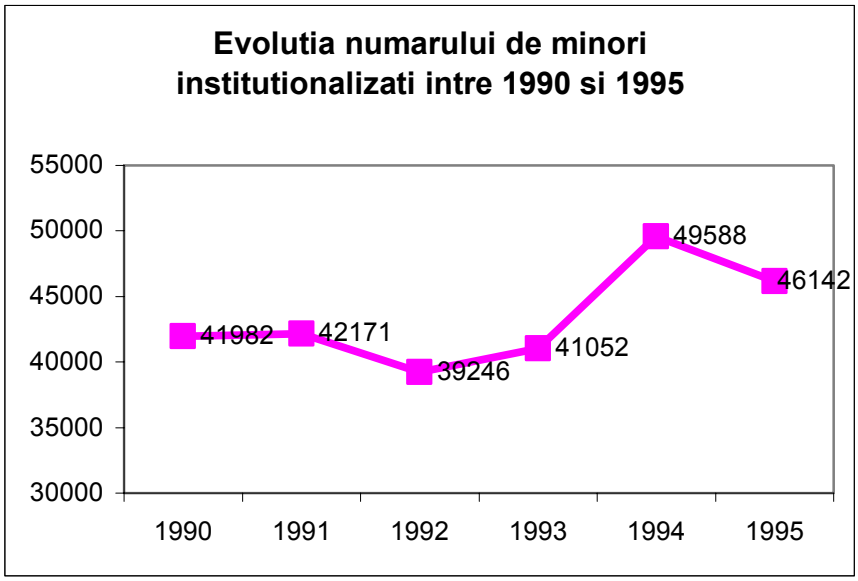
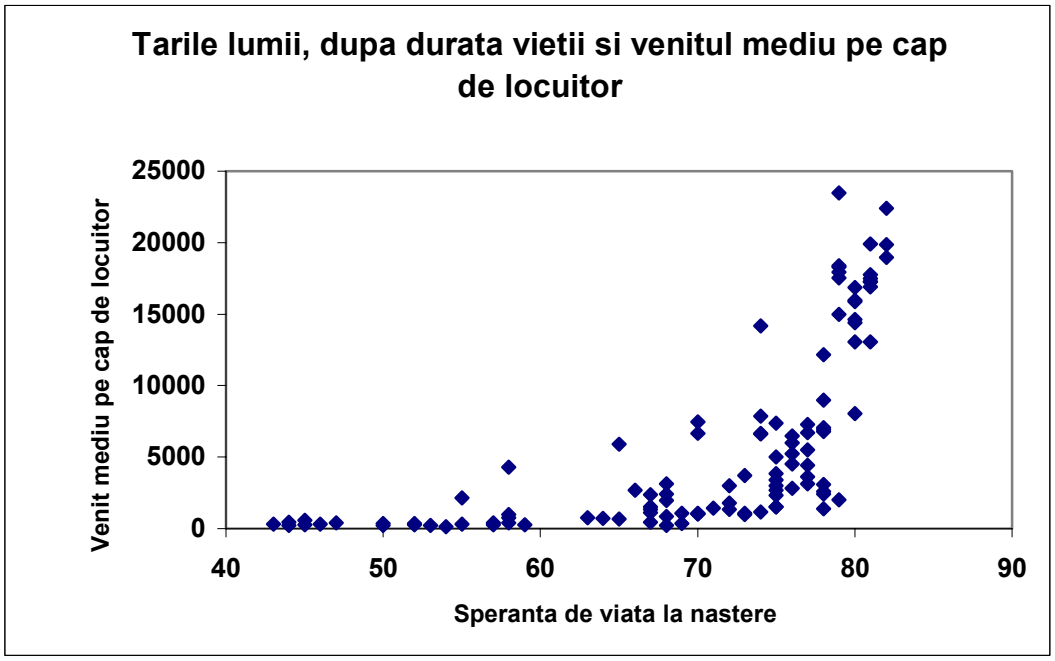


Figura 3.7 reprezintă un poligon de frecvențe etalând evoluția numărului de minori instituționalizați (frecvențe absolute) în cei 6 ani din perioada 1990-1995, în România.

3.4. Diagrama de împrăştiere

Diagramele de împrăştiere diferă față de celelalte tipuri de grafice anterior prezentate, prin aceea că reflectă rezultate individuale şi nu frecvențe realizate de valorile unei variabile la nivelul unei populații. Rezultatele observate pentru două variabile pentru diferiți indivizi sunt reprezentate grafic. Acest tip de prezentare grafică este adesea folosit pentru a sugera o posibilă asociere între două variabile de nivel interval sau rapoarte, sau pentru a sublinia în mod grafic existenta unei relații care a fost demonstrată prin folosirea unei analize statistice.

Figura 3.8. Exemplu de diagramă de împrăştiere



De exemplu, putem fi interesaţi în a examina legătura dintre speranţa de viaţă la naştere şi venitul mediu pe cap de locuitor pentru toate statele lumii, având date relative anului 1995 (vezi figura 3.8). Putem lansa ipoteza că speranţa de viaţă la naştere este mai mare cu cât venitul pe cap de locuitor este mai mare, şi invers. Din figura 3.8 putem vedea că multe ţări se situează aproape de axa orizontală, ceea ce denotă că realizează toate doar venituri mici, dar că din punctul de vedere al speranţei de viaţă ele prezintă variaţii considerabile, putând varia de la 40 ani până către 85 ani. Pe de altă parte, tot urmărind graficul, putem vedea că alte foarte multe ţări se plasează pe grafic într-o manieră verticală la capătul din dreapta al figurii. Aceasta înseamnă că ele au toate o speranţă de viaţă ridicată, apropiată de valorile maxime atinse pe glob, dar că din punctul de vedere al venitului mediu pe cap de locuitor sunt extrem de variate, de la cele mai sărace şi până la cele mai bogate. Totuşi nu putem spune că nu există nici o relaţie între cele două variabile, pentru că norul de puncte se structurează foarte clar după o curbă exponenţială, ceea ce denotă existenţa unei relaţii speciale între cele două variabile.

Câteodată, ca şi în exemplul nostru, variabilele respective pot fi reprezentate pe oricare din axele de coordonate, atâta timp cât nici una nu poate fi clar indicată ca variabilă independentă. Nu putem spună că speranţa de viaţă influenţează venitul mediu pe cap de locuitor, nici invers. Putem spune doar că ele sunt asociate, legate.

Frecvent, punctele reprezintă cazuri care pe diagrama de împrăştiere apar a fi distribuite în anumite forme (o linie dreaptă, sub formă de "U", sub formă de "J", etc.). Aceasta poate fi importantă pentru înţelegerea şi tragerea concluziilor din datele cercetării.

REZUMAT

În acest capitol am discutat diferite metode pentru reprezentarea grafică a distribuţiilor de frecvenţe sau pentru ilustrarea relaţiilor între două variabile. Graficele pot înlocui câteodată tabelele în prezentarea rezultatelor cercetării, sau se pot folosi şi împreună.

În selectarea unui grafic pentru includerea într-un raport, cercetătorul trebuie să aibă în vedere limitele şi specificul diferitelor grafice despre care s-a discutat. În unele cazuri se pot potrivi mai multe tipuri de grafice. Criteriul principal în alegerea graficului trebuie să fie claritatea şi corectitudinea prezentării în vederea obţinerii unui maximum de comunicare.

Întrebări pentru studiu

1. In graficele care folosesc valorile a două variabile, pe ce axă trebuie aşezate valorile variabilei dependente?
2. Prin ce diferă un grafic de bare de o histogramă?
3. Ce nivel de măsură trebuie să aibă variabila a cărei distribuție de frecvențe este reprezentată grafic printr-o diagramă de bare?
4. Dacă într-o agenție cu un buget anual de 36.000$ se alocă pentru cheltuieli de transport 9.000$, care este porțiunea dintr-o diagramă circulară care va corespunde segmentului transporturi?
5. De ce poligoanele de frecvențe descriu exact datele numai dacă ele sunt de nivel interval sau rapoarte?
6. Care dintre graficele cunoscute de dvs. poate sugera o posibilă relație între două variabile?

Exercițiu pentru acasă 3.

Construiți histograma distribuției de frecvențe de la Exercitiul pentru acasa 2.
Dați un titlu adecvat tabelului şi capului de tabel şi scrieți un mic comentariu care să valorifice cele mai relevante rezultate din tabelul de frecvențe (aproximativ 400-500 cuvinte).

Modul 4

Indicatori ai tendinței centrale

Reprezentările tabelare şi grafice ale informațiilor ne pot expune vizual rezultatele cercetărilor noastre. Apoi vine momentul când dorim să ne orientăm atenția către anumite caracteristici specifice datelor. În particular, am dori să rezumăm informația, să putem spune care este valoarea "tipică" , valoarea "normală" sau "tendința centrală".

Cuvântul *tipic* este unul obişnuit în vocabularul nostru. Vorbim de "clientul tipic" sau de "salariul normal pentru asistenții sociali începători", adesea fără a fi foarte exacți în înțelesul termenilor. Utilizarea *tipicului,* a *normalului, nivelului general, a ordinii de mărime* este o încercare de a găsi o singură valoare, sau o singură clasă caracteristică, care să descrie ce este reprezentativ pentru un întreg grup de numere sau clase.

În statistică există câteva modalități de găsire a valorii "tipice" pentru o variabilă. De aceea, folosirea unei terminologii precise este esențială. Orice termen trebuie folosit şi etichetat corect pentru a evita ambiguitățile şi confuziile. Grupăm diferitele descrieri ale răspunsurilor tipice găsite în datele noastre sub termenul de *tendință centrală*. Indicatorii tendinței centrale mai sunt cunoscuți în literatura de specialitate şi sub denumirea de *indicatori de poziție*. Fiecare indicator al tendinței centrale exprimă doar o cantitate limitată a informației despre un grup de valori, şi fiecare ne spune ceva diferit despre date. Pentru a înțelege diferența dintre diferiții indicatori ai tendinței centrale, trebuie să luăm în considerare modul de calcul al fiecăruia. Interpretarea indicatorului statistic presupune cunoaşterea modului în care a fost determinată valoarea lui.

Comparativ cu oricare alte categorii de indicatori, indicatorii tendinței centrale sunt cel mai des folosiți deoarece sunt uşor de calculat (găsit) şi sunt foarte utili. Ei au două întrebuințări de bază.

1. Furnizează o valoare unica, rezumativă, a datelor obținute. Ei reprezintă o încercare de a găsi o valoare (clasă sau număr), care să ne spună mai mult decât oricare altă valoare despre caracteristicile distribuției variabilei. De exemplu, o agenție care angajează câteva sute de asistenți sociali pe an poate da un salariu mediu lunar de stagiar de 150$. Acest singur număr ne ajută să rezumăm într-o singură valoare politica salarială a agenției pentru asistenții sociali începători.

2. Furnizează un numitor comun pentru compararea a două grupuri de informații. Dacă sunt obținute două numere - numărul mediu lunar al clienților care recurg la servicii familiale (Departamentul A) şi numărul mediu lunar al clienților care cer asistență financiară (Departamentul B) - se poate face o uşoară şi rapidă comparare a celor două departamente.

În acest capitol vom examina cei mai frecvenți indicatori ai tendinței centrale: modul, mediana şi media.

4.1. Modul

Modul este un indicator al tendinței centrale care ne spune care valoare din distribuția valorilor este observată cel mai frecvent.

Pentru date negrupate, ne referim la mod ca la o valoare care apare cel mai des. Dintre cele 10 valori prezentate mai jos, 7 este modul deoarece este întâlnit cel mai frecvent față de celelalte numere; numărul 7 se întâlneşte de trei ori.

2 4 5 6 7 7 7 8 8 9
mod

Pentru date grupate, ne referim la mod ca la punctul central al intervalului care conține frecvența cea mai mare. În Tabelul 4.1 avem informații legate de satisfacția profesională a unui număr de 50 asistenți sociali. Pentru aceste date grupate, intervalul care conține cea mai mare frecvență este 48-50, incluzând

numerele 48,49 şi 50. De vreme ce punctul de mijloc al acestui interval este 49, spunem că 49 *este valoarea modală* a distribuţiei, sau *modul*.

După unii autori, alegerea unui punct din *intervalul modal* care să reprezinte valoarea modală este nejustificată. După părerea lor, căutarea ar trebui să se oprească la simpla specificare a intervalului modal.

Tabelul 4.1 Distribuţia de frecvenţe cumulate pentru date grupate: satisfacţia în muncă
Rezultatele obţinute de asistenţii sociali de la agenţia XYZ

Rezultate	Frecvenţe absolute	Frecvenţe cumulate (Mari-Mici)	Frecvenţe cumulate (Mici-Mari)
36-38	2	50	2
39-41	2	48	4
42-44	4	46	8
45-47	1	42	9
48-50	7	41	16
51-53	3	34	19
54-56	4	31	23
57-59	1	27	24
60-62	1	26	25
63-65	4	25	29
66-68	5	21	34
69-71	1	16	35
72-74	6	15	41
75-77	5	9	46
78-80	1	4	47
81-83	3	3	50

Se poate întâmpla pentru datele negrupate ca mai mult decât o singură valoare să fie întâlnită cu aceeaşi frecvenţă. Dacă două sau mai multe valori au aceiaşi frecvenţă fiecare dintre aceste valori vor fi moduri ale datelor. Când exact două valori au aceiaşi frecvenţă, sau frecvenţe mult mai mari decât a celorlaltor valori, chiar dacă aceste două frecvenţe nu sunt strict egale, distribuţia este numită *bimodală*. Mai jos sunt prezentate opt valori care conţin două moduri.

4 5 5 6 6 7 8 9
modul 1 modul 2

Valorile 5 şi 6 sunt moduri şi se întâlnesc cel mai frecvent - de două ori fiecare. Dacă acelaşi lucru se întâlneşte în cazul datelor grupate, punctul de mijloc al intervalelor care conţin cel mai mare număr de cazuri sunt luate ca moduri ale datelor. Datele din Tabelul 4.2 arată o distribuţie bimodală. Modurile sunt punctele din mijlocul intervalelor care conţin cea mai mare frecvenţă. În acest caz modurile sunt 55 şi 52, mijloacele intervalelor 54-56, respectiv 51-53 ce conţin opt cazuri fiecare.

Restricţii de utilizare a modului

Dintre cei trei indicatori ai tendinţei centrale prezentaţi în această secţiune, modul este cel mai puţin restrictiv, dar şi cel mai slab. El se poate utiliza la variabile măsurate pe orice scală, inclusiv pentru cele de nivel nominal.

Avantajele şi dezavantajele modului

Modul ne dă o idee despre cea mai "populară" valoare dintr-o distribuţie, identificând valoarea cea mai des întâlnită. Dar modul nu este întotdeauna cel mai sugestiv indicator al tendinţei centrale deoarece îi

lipseşte precizia pe care ceilalţi indicatori din aceaşi categorie o posedă. Cea mai frecventă sau comună valoare a unei distribuţii nu este în mod necesar cea mai exactă descriere a valorii centrale. De exemplu, Tabelul 4.3 prezintă o distribuţiei de frecvenţe în care modul, nu este în mod evident în centrul distribuţiei, ci mai către sfârşitul ei (grupa 57-59). Pentru date de nivel ordinal, interval, rapoarte, putem cu uşurinţă obţine o descriere mai reprezentativă prin folosirea altor indicatori ai tendinţei centrale.

Tabelul 4.2 Distribuţia de frecvenţe cumulate pentru date grupate: satisfacţia în muncă
Rezultatele obţinute de personalul cleric de la agenţia XYZ

Rezultate	Frecvenţe absolute	Frecvenţe cumulate (Mari-Mici)	Frecvenţe cumulate (Mici-Mari)
45-47	4	31	4
48-50	7	27	11
51-53	8	20	19
54-56	8	12	27
57-59	4	4	31

Tabelul 4.3 Distribuţia de frecvenţe cumulate pentru date grupate: satisfacţia în muncă
Rezultatele obţinute de asistenţii sociali de la agenţia XYZ

Rezultate	Frecvenţe absolute	Frecvenţe cumulate (Mari-Mici)	Frecvenţe cumulate (Mici-Mari)
39-41	4	33	4
42-44	1	29	5
45-47	2	28	7
48-50	3	26	10
51-53	7	23	17
54-56	6	16	23
57-59	10	10	33

4.2. Mediana

Mediana este valoarea individului statistic (real sau virtual) care împarte populaţia în două submulţimi numeric egale, prima având valori mai mari sau egale cu acesta, iar cealaltă având valori mai mici sau egale cu ea.

În cele 11 valori prezentate mai jos, mediana este 15, deoarece 15 coincide cu punctul ce împarte valorile în două părţi egale, cinci valori în stânga lui 15 (10, 11, 12, 13, 14) şi tot cinci valori (16, 17, 18, 19, 20) în dreapta lui 15.

10 11 12 13 14 15 16 17 18 19 20
mediană

Primul pas în determinarea medianei unei distribuţii de valori este de a le aranja într-o ordine ca ce-a prezentată în Tabelul 2.2 din Modulul 2. Dacă numărul valorilor întâlnite este impar, mediana este definită ca punctul central al şirului de valori ordonate, iar dacă numărul lor este par mediana este media aritmetică a celor două valori din mijloc. Pentru cele şase valori de mai jos, 14,5 este mediana. Ea este media celor două valori centrale (14 şi 15).

12 13 14 15 16 17
mediana

Trebuie semnalat că aici mediana (14,5) nu coincide cu nici o valoare. Această situație pune în evidență faptul că - contrar unei concepții greșit înțelese - mediana nu este doar o simplă valoare a cazului mijlociu. Câteodată aceasta coincide cu valoarea unui caz mijlociu, dar în exemplul anterior se vede că poate fi doar o medie a valorilor cazurilor mijlocii. Aceasta se întâmplă deoarece mediana adevărată (punctul de mijloc al unei serii) este influențată de o serie de factori ca: numărul par sau impar al cazurilor, frecvența mai mare decât unu a cazurilor cu valori apropiate de centrul seriei; valorile cu frecvența zero, etc. Mediana se obține în mod obișnuit folosind o procedură specială desemnată pentru acest scop (pe care nu o tratăm aici). Pentru detalii se poate consulta, de exemplu, volumul "Metode statistice aplicate în ştiințele sociale", coord.T.Rotariu, Editura Polirom, 1999.

Restricții de utilizare a medianei

Deoarece mediana împarte distribuția valorilor în două părți egale, este necesar ca valorile să poată fi aranjate într-o ordine. Din acest motiv, este necesar ca datele să fie măsurate pe o scală cel puțin la nivel ordinal.

Avantajele şi dezavantajele medianei

Dintre cei trei indicatori ai tendinței centrale tratați în această secțiune, mediana este cea mai stabilă, ea este mai puțin afectată de valorile extreme întâlnite în distribuție. Cele două seturi de câte treisprezece valori care urmează au aceiaşi valoare mediană, dar observați valorile lor extreme foarte diferite.

Distribuția A

1	14	15	16	17	18	19	20	21	22	23	24	50
						↑						
						mediană						

Distribuția B

13	14	15	16	17	18	19	20	21	22	23	24	25
						↑						
						mediană						

În Distribuția A valorile extreme, sunt mult diferite față de celelalte valori. În distribuția B, valorile extreme sunt 13 şi 25, şi ele sunt mult mai apropiate de celelalte valori ale distribuției. De fapt ambele distribuții au aceiaşi mediană 19; aceasta demonstrează că valorile extreme ce se întâlnesc la începutul şi la sfârşitul distribuției nu au efect asupra medianei sau au un efect mic. Această calitate, stabilitatea, face din mediană, un indicator mai des utilizat pentru măsurarea tendinței centrale.

Prin comparație; modul poate fluctua mai mult decât mediana. Modul prezintă doar cea mai des întâlnită valoare într-o distribuție de valori şi nu apare în mod necesar în centrul distribuției.

Dezavantajul major al medianei este acela că obținerea ei este mai dificilă decât calculul modului prezentat anterior. Dublurile valorilor sau alți factori, necesită frecvent utilizarea unei formule mai complicate. Din fericire în cazul analizării mai multor date, ne putem folosi de calculator.

În calculul medianei se face abstracție de jumătate din valorile seriei, de acea se consideră că suferă de lipsă de precizie, dezavantaj care poate fi compensat prin utilizarea mediei, un alt indicator al tendinței centrale.

4.3. Media

Măsurarea tendinței centrale cu un indicator mai sofisticat decât modul sau mediana este posibilă atunci când informația este de nivel interval sau rapoarte. Dacă o variabilă poate fi considerată de nivel interval sau de rapoarte, utilizarea mediei ca indicator al tendinței centrale trebuie luată în considerare.

Deoarece media este uşor de înțeles, şi deci mai folositoare, ea este adesea utilizată ca măsură a tendinței centrale. În sens matematic, *media reprezintă suma tuturor valorilor variabilei împărțită la numărul de valori*. Media poate fi definită şi în alte moduri - care pot fi luate ca tot atâtea proprietăți ale ei - ca de exemplu, *valoarea comună pe care ar trebui să o ia toți indivizii din populație pentru ca suma valorilor să nu se schimbe*. Media se poate calcula cu formula:

$$\text{Media} = \frac{\text{Suma tuturor valorilor individuale}}{\text{Volumul populației statistice}}$$

Pentru date negrupate, ca în cazul valorilor de mai jos, putem obține media prin însumarea valorilor şi împărțirea totalului la numărul valorilor.

4 4 5 7 10

Nu avem nevoie să le aranjăm în ordine ca în cazul calculării medianei. Pentru aceste date este necesară doar efectuarea operațiilor.

Media = (4+4+5+7+10) / 5
= 30 / 5
= 6 (media)

Media se calculează uşor şi cu ajutorul tabelelor de frecvențe, utilizându-se la numărător suma produselor dintre valori şi frecvențe. Dacă tabelul de frecvențe este cu intervale de grupare, în calcule se folosesc valorile centrale ale intervalelor.

EXERCITIU: Calculați indicatorii tendinței centrale pentru numărul de membrii ai familiilor beneficiari de ajutor social din Tabelul 5.8.

Restricții de utilizare a mediei

Pentru a folosi media, datele trebuie să fie cel puțin de nivel interval. Calcularea mediei pentru nivelul nominal sau ordinal nu are sens. De exemplu, nu pentru orice şir de valori este potrivit a calcula media. Un student poate fi al treilea clasat în primul an de studii, al patrulea în cel de al doilea, şi al doilea în următorul. Suma rangurilor şirului parcurs este 3+4+2, deci 9 şi astfel "media şirului este 3 (9/3 = 3) dar nu are nici un sens să vorbim de acesta. Nivelul interval pentru date arată că există o distanță egală între intervale sau nivele. Doar în acest ultim caz calcularea mediei este valabilă. Când scala folosită nu este la nivelul de măsurare interval media nu este utilizabilă.

O a doua cerință pentru folosirea mediei este legată de volumul eşantionului. Deoarece valorile extrem de mari sau valorile extrem de mici pot distorsiona media foarte uşor, este mai bine ca ea să fie utilizată în eşantioane relativ mari în care valorile aberante să cauzeze mai puține distorsiuni. Pentru ca media să fie utilizată ca indicator unic al tendinței centrale, mai este necesar ca forma distribuției să fie simetrică sau "normală". Vom explora conceptul de distribuție normală în Modulul 6.

Avantajele şi dezavantajele mediei

Media este frecvent folosită şi mulți dintre noi suntem familiarizați cu ea. Media implică mai multă precizie decât modul sau mediana. Pentru calculul ei se folosesc valorile precise ale *fiecărui* rezultat particular (nu doar unele din valori sau frecvențe). Această caracteristică poate promova claritate sau distorsiune, depinzând de cât de asemănătoare sunt rezultatele între ele.

Un mare dezavantaj al mediei este că poate fi folosită doar pentru date de nivel interval sau rapoarte. Această regulă limitează posibilitatea folosirii acesteia pentru date măsurate la nivel nominal sau ordinal. Acest dezavantaj este resimțit mai ales în cercetarea din asistența socială. Există multe împrejurări în cercetările noastre, unde media pur şi simplu nu poate fi folosită. De exemplu, multe instrumente de măsurare a atitudinii, generează un număr ordinal pentru fiecare caz, dar doar atâta nu este suficient pentru a trata informațiile ca fiind de nivel interval sau rapoarte. Datorită popularității mediei, unii insistă în mod eronat în aplicarea ei şi în astfel de situații. Este mai recomandabil să se folosească mediana în cazul în care datele nu pot fi justificate ca fiind de nivel interval sau rapoarte.

Calitatea mediei de a fi un indicator precis, senzitiv la orice schimbare, se poate transforma, în unele situații, în dezavantaj. Senzitivitatea poate fi considerată excesivă atunci când schimbările se petrec la extremele intervalelor de variație. Putem ilustra acest dezavantaj cu media de vârstă a clienților. Presupunem că avem zece clienți cu următoarele vârste:

25 26 27 28 30 31 32 32 33 **76**

Media rezultată din însumarea valorilor vârstelor, împărțite la numărul de clienți, 340 / 10, este 34. Totodată nouă clienți au 33 de ani sau mai puțin. În acest caz este evident ca o singură persoană, cea cu vârsta 76, este principalul responsabil de distorsionarea "mediei de vârstă a grupului". In acest caz valoarea de 76 măreşte considerabil media de vârstă a grupului. Treizeci şi patru nu reflectă exact o măsură a tendinței centrale. Modul şi mediana pentru aceste date este de 32 şi respectiv 30,5; valori cu siguranță mult mai adecvate, în sensul de a fi mai centrale, mai tipice. Acest exemplu sugerează că atunci când distribuția conține câteva valori deviante (marcant depărtate față de celelalte) mediana este, în mod obişnuit, preferabilă mediei. In cazul tabelelor de frecvente, observând poligonul frecvențelor, vom vedea că acesta este foarte asimetric. Cel mai bine este raportam ambii indicatori ai tendinței centrale, atât media cât şi mediana, sau chiar toate cele trei (plus modul) – pentru ca astfel vom produce o imagine mai bună asupra distribuției.

4.4. Modul, mediana sau media ?

La întrebarea *care dintre indicatorii tendinței centrale trebuie folosit?* nu este întotdeauna aşa de simplu de răspuns. In timp ce avem câteva reguli generale de aplicare sigure, decizia finală este adesea o rezolvare etică. Peste toate acestea, cercetătorii doresc să folosească indicatorii tendinței centrale pentru a furniza cititorilor o imagine mentală clară, o descriere scurtă a ceea ce arată datele. Sunt câteva situații în care nici o valoare a variabilei nu este tipică şi folosirea oricărui indicator al tendinței centrale poate induce în eroare la fel de bine. Un exemplu care foloseşte cazuri din agenția XYZ va ilustra acest punct.

Tabelul 4.4 descrie o distribuție de frecvențe cu numărul interviurilor pentru toate cazurile închise în luna decembrie. Variabila "numărul interviurilor" este de nivelul rapoarte, iar numărul mare de cazuri sugerează că media poate fi cea mai bună măsură a tendinței centrale. Cu toate că, aşa cum am arătat mai devreme, media necesită o distribuție care este în general simetrică, forma poligonului de frecvențe, în formă de U, este primul avertisment, acela că media, sau de fapt doar ea singură, nu este o alegere corespunzătoare. Numărul mediu al interviurilor din Tabelul 4.4 este 4,56 (numărul total de interviuri este de 1345, iar rezultatul împărțirii ei la 295 - câți clienți a avut agenția -, este 4,56). Deci, în medie fiecărui client îi revin 4,56 interviuri.

Totuşi intervievarea unui client de patru sau de cinci ori, nu este tipică pentru clienții agenției. Există patru valori (1, 2, 3 şi 10) care se întâlnesc la fel de frecvent, sau mai frecvent decât 4 sau 5. Din cauză că primele trei valori (1, 2 şi 3) sunt toate grupate în stânga mediei, media apare ca fiind prea mare ca să reflecte ceea ce este tipic.

Tabelul 4.4 Distribuție de frecvențe: Număr de interviuri pentru cazurile închise în luna decembrie la agenția XYZ

Număr de şedințe	Frecvențe absolute (număr de cazuri închise)
1	55
2	35
3	55
4	30
5	25
6	15
7	10
8	20
9	10
10	40
Total cazuri	295

Dacă media nu poate descrie reprezentativ informația, ce se poate spune despre mediană? Ea cade pe 4 şi e foarte aproape de valoarea cea mai comună (3). Ca unică alegere pentru o măsură a tendinței

centrale, este suficient de bună. Dar ea nu este sugestivă pentru grupul clienților care au fost intervievați de zece ori (40 de cazuri), un fapt care poate fi un surprinzător şi demn de semnalat pentru o agenție care în general consideră să se angajeze în intervenții în criză, pe termen scurt. De asemenea nu confirmă constatarea predictibilă că un număr mare de clienți (55) au fost intervievați o singură dată. Pe scurt, mediana poate fi mai bună decât media sau modul pentru o prezentare a tendinței centrale dar este departe de perfecțiune pentru această informație.

Distribuția este bimodală, cu două moduri care cad pe unu şi trei interviuri. Dar dacă am folosi doar modul, am putea sugera că "unu" sau "trei" interviuri (ambele numere mici) sunt tipice pentru cazuri, dar de fapt mai puțin de jumătate din toate cazurile au fost intervievate de patru ori sau mai puțin. Ca şi mediana, modul singur nu sugerează posibilitatea că "intervențiile în criză" se repetă adesea într-un număr mare de şedințe. Aşa cum am subliniat, modul este mai adecvat nivelului nominal. El tratează diferitele valori ale variabilei ca şi când între ele există doar diferențe de clasă şi nu ține cont (aşa cum s-a întâmplat şi în exemplul nostru) dacă există diferențe cantitative reale.

Câteodată nici un indicator al tendinței centrale nu este ideal pentru a descrie ceea ce este tipic pentru datele concrete. Exemplul nostru ilustrează o situație în care oricare din indicatori are un potențial de eroare. Aici toți trei - modul, mediana sau media - pot contribui într-un fel propriu la interpretarea datelor. Faptul că distribuția este bimodală, cu modurile sunt la unu şi la trei interviuri, indică faptul că perioadele scurte rămân o caracteristică obişnuită în agenție. Aici mediana reflectă cel mai bine ceea ce este nivelul general. Ea se bucură de o oarecare precizie, mai mare decât modul, dar mai mică decât media şi sugerează că perioadele scurte de tratament nu sunt chiar aşa de tipice pentru agenție cum am presupus la o primă vedere. Dacă se prezintă şi media, valoarea ei depăşind 4 interviuri, se aduce o dovadă puternică a faptului că un număr mare de clienți au fost intervievați mult mai frecvent decât ne-am aşteptat.

Exemplul care foloseşte informația din Tabelul 4.4 este un argument în recomandarea de a lua în considerare şi de a raporta mai mulți indicatori ai tendinței centrale. Un cititor experimentat al rapoartelor de cercetare, sau chiar şi unul care înțelege corect modul, mediana si media, trebuie să fie capabil să le compare, ele oferindu-i o bună imagine asupra datelor deținute. Oricare dintre ei luat singur poate deforma realitatea; toți trei împreună îi vor oferi o imagine plină de acurateție - ceea ce este, de fapt, un obiectiv al statisticii descriptive.

REZUMAT

În acest capitol, am discutat diferiți indicatori ai tendinței centrale şi modurile prin care se pot folosi pentru a prezenta valoarea tipică, normală, pentru un set de date. Trebuie să subliniem că anumiți factori cum sunt nivelul de măsurare, volumul eşantionului şi prezența sau absența oricărui rezultat deviant, toate trebuie luate în considerare în selecția unuia sau a mai multor indicatori ai tendinței centrale.

Indicatorii de poziție, cum se mai numesc indicatorii tendinței centrale, sunt un fel de ajutor şi nu un substituent pentru înțelegerea variațiilor individuale în interiorul unui set de date. Măsura tendinței centrale este o cale simplă şi rapidă de comunicare a caracteristicilor subiecților când se doreşte o precizie minimă. Acelaşi lucru se poate spune şi despre prezentarea tabelară sau despre prezentarea grafică a datelor, etica cercetării cerând ca în alegerea indicatorilor să aibă prioritate onestitatea comunicării.

Întrebări pentru studiu

1. De ce modul este cel mai potrivit pentru a fi folosit pentru datele de nivel nominal?
2. Cum arată un poligon de frecvențe când distribuția poate fi descrisă ca bimodală?
3. De ce este mediana mult mai bună şi mai precisă în descrierea informației de nivel ordinar, de cât modul?
4. In ce situații media poate fi nepotrivită pentru nivelul general al notelor într-o grupă de 12 studenți?
5. De ce considerăm media a fi cel mai precis indicator al tendinței centrale față de mediană şi mod?
6. Cum se explică faptul că câteva valori extreme mai ridicate sau mai scăzute ale variabilei tind să distorsioneze media?
7. Dacă nu sunt îndeplinite toate cerințele necesare pentru a folosi media, care este alegerea alternativă pentru a măsura tendința centrală?
8. De ce folosirea unui singur indicator al tendinței centrale poate fi inadecvată descrierii unui set de date?

9. În ce situație folosirea mai multor indicatori ai tendinței centrale furnizează o mai bună descriere a valorilor variabilei, decât unul singur?

Exercițiu pentru acasă 4.

Un asistent social implicat într-un program de terapie familială lucrează cu 10 familii, care au următorul număr de copii minori în întreținere: 1, 0, 1, 2, 1, 3, 2, 2, 2, 2

a. Precizați individul statistic, variabila şi nivelul scalei de măsură.
b. Calculați media, mediana, modul.
c. Folosiți indicatorii obținuți mai sus pentru a descrie situația copiilor din grupul celor 10 familii aflate in evidenta asistentul social

Modul 5

Indicatori ai împrăştierii

In modulul anterior am prezentat trei indicatori ai tendinței centrale. Dacă am opri aici analiza noastră asupra informației, descrierea oricăror distribuții de valori ar fi trunchiată sau greşit înțeleasă. Modulul anterior a descris punctele în jurul cărora toate valorile distribuției înclină să se focalizeze, dar nu explică modul în care valorile au fost distribuite în jurul acestor puncte.

Distribuțiile A şi B par a avea aceiaşi medie: 3. Totuşi modul în care fiecare set de valori este distribuit în jurul mediei este diferit.

Distribuția A:	0	2	3	4	6
Distribuția B:	2	3	3	3	4

In general vorbind despre descrierea acestor două distribuții de câte cinci valori fiecare, putem spune că valorile Distribuției B sunt grupate mai apropiat în jurul mediei 3, în timp ce valorile Distribuției A sunt mult mai larg împrăştiate în jurul aceleiaşi medii.

Felul în care se împrăştie valorile în jurul indicatorilor tendinței centrale al unei distribuții, este prezentat prin ceea ce numim *indicatori de împrăştiere*. Alte denumiri consacrate sunt *indicatori ai variabilității* sau *indicatori de dispersie*.

Ei urmăresc, în esență, să măsoare gradul de împrăştiere a indivizilor în cadrul seriei de valori pe care aceştia le iau. Tradusă problema în alți termeni, indicatorii de împrăştiere caracterizează o populație statistică din punctul de vedere al *omogenității / eterogenității*, în raport cu o variabilă dată. În anumite situații, în special când variabila reflectă anumite ierarhii socialmente recunoscute sau valorizate între indivizii statistici (adică în sensul că valori mai mari sunt considerate ca fiind mai bune sau preferabile celor mici, cum e cazul salariilor de pildă), indicatorii de împrăştiere reflectă *gradul de inegalitate între indivizi*, în raport cu factorul social tradus în caracteristică.

În secțiunile statisticii consacrate aşa-ziselor "modele explicative", indicatorii de dispersie capătă un rol esențial prin faptul că li se atribuie semnificația de a reproduce *gradul de nedeterminare*, de variabilitate a unui fenomen. Or, explicația, cel puțin din punct de vedere statistic, înseamnă reducerea acestei nedeterminări sau variabilități, adică sporirea posibilității de a prezice starea unui fenomen. Într-adevăr, aşa cum am sugerat deja, un fenomen este tradus într-o caracteristică de tip statistic, care, la rândul ei este exprimată printr-un indicator al tendinței centrale (să spunem media). A afirma că, de pildă, media notelor la matematică într-o clasă de elevi este de 8,50 înseamnă a afla ceva despre nivelul de cunoştințe în domeniu al şcolarilor în cauză. Dispersia notelor din care rezultă această valoare medie va indica măsura în care valoarea individuală este reprodusă prin cea de grup sau calitatea predicției situației indivizilor prin mărimea la nivelul colectiv. Dacă în analiză se introduce o variabilă nouă, cu rol explicativ pentru cea inițială, atunci, presupunând cunoscută poziția indivizilor după această caracteristică socotită independentă, se va reduce nedeterminarea situației lor după variabila de bază, în măsura în care între cei doi factori există o legătură. Altfel spus, indicatorul de variabilitate va pune în evidență cât din nedeterminarea variabilei cercetate se poate reduce prin introducerea în modelul explicativ a unei noi variabile.

În continuare vom prezenta cinci indicatori mai cunoscuți ai variabilității: amplitudinea, abaterea intercuartilă, abaterea medie, varianța şi abaterea standard.

5.1. Amplitudinea

Amplitudinea este un indicator al variabilității care determină distanța dintre valoarea maximă şi valoarea minimă dintr-o distribuție. Putem exprima aceasta prin formula:

Amplitudinea = Valoarea Maximă - Valoarea Minima

Considerăm distribuția C si D de mai jos

Distribuția C:	1	5	5	5	5
Distribuția D:	1	5	5	5	**9**

Amplitudinea distribuției C se calculează astfel:
Amplitudinea = 5 -1
= 4 (Amplitudinea)

Amplitudinea distribuției D se calculează astfel:
Amplitudinea = 9 - 1
= 8 (Amplitudinea)

Chiar dacă distribuțiile C şi D au acelaşi număr de valori, intervale de valori asemănătoare, totuşi amplitudinea mai mare a distribuției D indică o mai mare variație a valorilor acestei distribuții.

Cerințe pentru amplitudine

Amplitudinea poate fi folosită în analizele statistice doar când informația este de *nivel interval sau rapoarte*. Această cerință este frecvent încălcată atunci când amplitudinea se foloseşte pentru distribuțiile datelor măsurate la nivel ordinal. Amplitudinea presupune existența distanțelor egale între toate valorile dintr-o distribuție. Aşa cum am subliniat în Modulul 1, pentru datele de nivel ordinar nu ştim cu siguranță cât este de mare distanța dintre doua valori oarecare. Ştim cu siguranță, doar că o valoare este mai mare (sau mai mică) decât o altă valoare, dar nu şi cu cât este mai mare (mai mică).

Avantaje şi dezavantaje ale amplitudinii

Amplitudinea este o măsură utilă a variației, ea poate fi calculată repede şi uşor. Odată ce am identificat valoarea limită cea mai înaltă şi valoarea limită cea mai scăzută a unui set de valori, putem construi mult mai uşor distribuția frecvențelor. Calcularea amplitudinii ne permite să determinăm numărul intervalelor folosite pentru gruparea datelor şi mărimea cea mai potrivită pentru intervale.

Amplitudinea este un indicator instabil. Ea este calculată pe baza valorii maxime şi minime dintr-o distribuție, deci o singură valoare deviantă poate mări considerabil amplitudinea. De exemplu, să considerăm distribuțiile E şi F de mai jos

Distribuția E:	10	11	12	13	14	15	16	17	**90**
Distribuția F:	10	11	12	13	14	15	16	17	18

În distribuția E, amplitudinea este 80 (90-10=80). În distribuția F, amplitudinea este 8 (18-10=8). Diferența între amplitudinile 80 şi 8, este corectă şi este cauzată de valoarea extremă 90 din distribuția E. Această proprietate este una nedorită. Aceasta face ca din compararea amplitudinilor a două sau mai multe distribuții, să rezulte posibile erori.

5.2. Abaterea intercuartilă

Cuartilele sunt acele valori ale indivizilor statistici (reali sau virtuali), din ierarhia valorilor de la cea mai mică la cea mai mare, care împart mulțimea în patru părți egale. Similar se pot utiliza *decilele*, care impart populația în 10 părți, *centilele* sau *percentilele* care o împart în 100 părți, sau, şi mai general, *cuantilele* care o împart într-un număr oarecare de părți de volume egale.

Tabelul 5.1 Valori şi centilele lor

Valori		Centile
1		
2		
3		
cuartila 1	=	3,5
4		
5		
6		
cuartila 2	=	6,5
7		
8		
9		
cuartila 3	=	9,5
10		
11		
12		

În loc de a măsura distanța dintre maximul şi minimul valorilor pentru a obține amplitudinea, putem măsura distanța dintre cea de a treia cuartilă (sau centila 75%) şi prima cuartilă (sau centila 25%). Această distanță este cunoscută ca abaterea intercuartilă. Ca şi indicator al variabilității, este mult mai stabilă decât amplitudinea pentru că, acele câteva valori extreme mai mari sau mai mici nu îl pot modifica, aşa cum ar face în cazul calculării amplitudinii. Dacă cuartila 3 şi cuartila 1 sunt deja determinate (prin calcul sau prin estimare într-un grafic cumulativ), abaterea intercuartilă se calculează prin scădere directă. Îl putem exprima prin formula următoare.

Abaterea intercuartilă = cuartila 3 – cuartila 1

Considerăm cele 12 valori din Tabelul 5.1. Cuartila 1 cade între valorile 3 şi 4, şi cuartila 3 cade între valorile 9 şi 10. Abaterea intercuartilă al datelor prezentate în Tabelul 5.1 este:

Abaterea intercuartilă = 9,5 - 3,5
= 6 (abaterea intercuartilă)

Abaterea intercuartilă se foloseşte in situații similare cu mediana. Ca şi mediana ea este folositoare în distribuții de nivel interval sau rapoarte unde câteva valori extreme pot distorsiona valoarea unor indicatori ai variabilității, cum ar fi abaterea standard (ce urmează a fi prezentată).

Cerințe pentru abaterea intercuartilă

Singura cerință necesară pentru calcularea abaterii intercuartile este să avem informație de nivel interval sau rapoarte.

Avantajele şi dezavantajele abaterii intercuartile

Principalul avantaj al abaterii intercuartile este că are proprietatea de a fi un indicator al variabilității mult mai stabil decât amplitudinea. El nu este influențat de valorile extreme ale distribuției. Este necesar un timp mai mare pentru a-l calcula decât timpul necesar pentru aflarea amplitudinii.

Abaterea semi-intercuartilă

Abaterea semi-intercuartilă este jumătate din abaterea intercuartilă. Putem scrie aceasta sub formă de formulă:

Abaterea semi-intercuartilă = (centila 75% - centila 25%) / 2

Pentru distribuția valorilor din Tabelul 5.1 abaterea semi-intercuartilă este:

Abaterea semi-intercuartilă = (9,5 - 3,5) / 2
= 3 (abaterea semi-intercuartilă)

5.3. Abaterea medie

Amplitudinea, abaterea intercuartilă şi abaterea semi-intercuartilă, au fiecare avantaje şi sunt folositoare în anumite situații, dar nici una dintre cele trei nu utilizează *fiecare* valoare a distribuției pentru determinarea ei statistică (deci suferă de lipsă de senzitivitate). Când dorim un indicator al variabilității care să ia în considerare *toate* valorile (sau toate observațiile) putem alege abaterea medie.

În loc să privim doar la cele două valori extreme dintr-o distribuție, ne putem pune problema descrierii variabilității prin indicarea gradului în care toate valorile grupului diferă (deviază față) de media distribuției. Cu cât valorile înclină să fie mai împrăştiate față de medie, cu atât ele vor tinde să difere unele față de altele. Media grupului este un punct de referință ideal din punct de vedere matematic, dar suma abaterilor față de medie este întotdeauna zero (ceea ce înseamnă că media este, în fapt, centrul de greutate al distribuției). Această proprietate este foarte importantă deoarece ne asigură că valoarea variabilității statistice va reflecta doar diferențele dintre membrii grupului şi nu dintre valorile lor absolute.

Tabelul 5.2 Abaterile față de medie

Valoarea	-	Media	=	Abaterea față de medie
1	-	3	=	-2
2	-	3	=	-1
3	-	3	=	0
4	-	3	=	1
5	-	3	=	2
Total				0

Tabelul 5.2 listează cinci valori, media lor şi mărimea abaterilor. Mărimea abaterii pentru o valoare oarecare este diferența dintre valoarea observată şi media grupului.

Abaterea pentru o valoare = valoarea - media

Pentru a calcula abaterea medie pentru datele din Tabelul 5.2 procedăm după cum urmează:

$$\text{Abaterea medie} = \frac{\text{Suma abaterilor valorilor (ignorând semnul + sau -)}}{\text{Numărul de valori din distribuție}}$$

$$= \frac{(2) + (1) + (0) + (1) + (2)}{5}$$

$$= 6 / 5$$

$$= 1{,}25 \text{ (abaterea medie)}$$

Valoarea de 1,25 este media abaterilor valorilor față de media lor (din Tabelul 5.2). Majoritatea oamenilor înțeleg uşor conceptul distanței medii față de media unei distribuții. Este relativ uşor de calculat şi de interpretat.

Abaterea medie poate fi calculată folosind şi alte valori de referință, nu numai media. De exemplu abaterea medie față de o valoare oarecare, să zicem "a". Se poate demonstra că valoarea de referință pentru care abaterea medie este cea mai mică este mediana.

În ciuda calităților şi expresivității abaterii medii, ea se foloseşte rareori în practică, pentru că se preferă un alt indicator de împrăştiere: abaterea standard. Şi acesta este un indicator cu expresivitate mare şi în plus, este folosit mult în calculul testelor statistice şi în analiza explicativă.

5.4. Abaterea standard şi varianța

Abaterea standard (sau deviația standard) este un indicator al dispersiei larg utilizat. Pătratul ei, cunoscut sub numele de *varianță*, este o componentă importantă a altor formule care examinează relațiile dintre variabile. La fel ca şi abaterea medie, pentru calcularea sa, abaterea standard ia în considerare valorile tuturor cazurilor. Dar nu se ignoră pur şi simplu semnul diferenței dintre fiecare mărime şi medie, aşa cum se făcea la abaterea medie, ci se utilizează ridicarea la pătrat pentru a scăpa de valorile negative.

În timp ce abaterea medie ne poate spune care este ordinul de mărime a abaterilor valorilor față de medie, abaterea standard face mai mult. Folosită împreună cu media în situații corespunzătoare, ne permite (1) să avem o imagine completă asupra locului unde se plasează o mărime față de celelalte valori, (2) să ştim ce procentaj de valori dintr-o distribuție se plasează în interiorul unei distanțe specificate față de medie şi (3) să reconstruim distribuția tuturor valorilor unei variabile. Utilizările (2) şi (3) sunt valabile numai în cazul în care distribuția de frecvențe a datelor empirice se apropie de modelul matematic al distribuției normale.

Cerințe pentru abaterea standard

Abaterea standard cere date de nivel interval sau rapoarte. Este de asemenea mult mai potrivit să fie folosită pentru eşantioane suficient de mari şi cu variabile care pot fi reprezentate grafic sub forma unui poligon de frecvențe simetric, de o formă particulară de clopot, numită şi distribuție normală (discutată în detaliu în Modulul 6).

Când se utilizează abaterea standard

Media sugerează care este nivelul general al datelor pentru o variabilă, dar ea nu ne spune cum sunt dispersate celelalte valori în jurul mediei. Această informație ne este dată de către abaterea standard.

De exemplu, putem avea două distribuții care au medii identice şi acelaşi număr de valori), dar variabilități diferite, ca în distribuțiile G şi H.

Distribuția G:	8	9	10	10	13
Distribuția H:	1	5	10	16	18

Figurile 5.1 şi 5.2 reprezintă câte cinci valori pentru fiecare din cele două distribuții. Figurile 5.1 şi 5.2 înfățişează valorile celor două distribuții ca şi greutăți pe o bară gradată. Putem observa cu uşurință din Figurile 5.1 şi 5.2 că distribuția H este mult mai variabilă decât distribuția G. Câteodată nici acest tip de reprezentare nu poate vizualiza care dintre cele două distribuții este mai variabilă.

Figura 5.1 Distribuția de frecvențe G

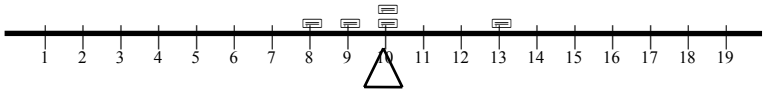
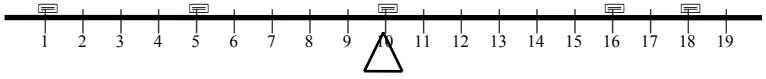


Figura 5.2 Distribuția de frecvențe H



Considerăm distribuțiile I şi J:

Distribuția I:	1	4	10	11	14
Distribuția J:	2	5	8	12	13

Reprezentând grafic valorile celor două distribuții (ca şi greutăți pe o bară gradată) pentru a determina care este mai variabilă, poate fi până la urmă o simplă pierdere de vreme (ca la figurile 5.3 şi 5.4). In acest punct avem într-adevăr nevoie de un indicator cum este abaterea standard.

Figura 5.3 Distribuția de frecvențe I

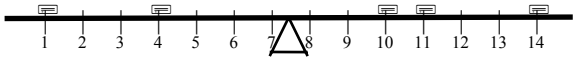
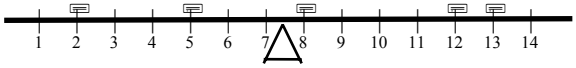


Figura 5.4 Distribuția de frecvențe J



Sa urmarim si un exemplu mai complex, cu date sistematizate intr-un tabel de frecvente. Mărimea mediei pe scala depresiei pentru un anumit grup de clienți care primesc un tratament de grup de tip nou (grupul A) poate fi la fel ca pentru ceilalți clienți care primesc un tratament de grup obişnuit, dar nu putem şti, şi nici nu putem presupune, că distribuția totală a valorilor ar fi tot aceiaşi. Tabelul 5.3 ne prezintă valorile măsurate pe o scală a depresiei pentru 20 de clienți din Grupul A față de valorile pentru toți ceilalți clienți care primesc tratamentul de grup pentru depresie. Ei au o medie identică de 74 puncte.

Tabelul 5.3 Distribuția de frecvențe a valorilor depresiei clienților unei agenții

Mărimea depresiei	Frecvențe Grupul A	Frecvențe Ceilalți clienți tratați în grup
50	0	10
60	4	10
70	8	25
80	4	45
90	4	5
100	0	5
Total	20	100

Observăm că valorile depresiei pentru Grupul A includ valori de la 60 la 90, pe câtă vreme pentru cei din alte grupuri de tratament distribuția include valori mai joase, ca 50, şi mai înalte, ca 100. Putem cu uşurință observa că Grupul A are amplitudinea de 30 puncte (90-60=30), pe câtă vreme celelalte grupuri de tratament au amplitudinea de 50 de puncte (100-50=50). Valoarea medie a depresiei pentru ambele categorii de clienti este aceeasi, 74 de puncte. Amplitudinea insa este o valoare prea grosiera a diferentelor de imprastiere. Abaterea standard ia in considerare toate valorile celor doua grupuri si tine cont si de volumul diferit al celor doua populatii (20 persoane, fata de 100 persoane). Abaterile standard ale celor doua grupuri, 10 puncte pe scala depresiei pentru grupul A si 12 puncte pentru grupul celorlalti, arata ca grupul A este mai omogen dar nu intr-atat de diferit de omogenitatea grupului celorlati pe cat ar parea sa induca indicatorul de amplitudine (50 este aprope dublu fata de 30).

5.5. Algoritmul de calcul al abaterii standard

Pentru a decide care distribuție este mai variabilă, trebuie să calculăm abaterea standard pentru amândouă distribuții. După aceea le vom compara: cea mai mare abatere standard va indica cea mai mare variabilitate în distribuție. Abaterea standard poate fi imaginată ca o distanță dintre medie şi un anumit punct de pe poligonul frecvențelor unei distribuții. Compararea a două abateri standard calculate indică care distribuție are acest punct mai depărtat de medie (aceea care are abaterea standard cea mai mare).

Abaterea standard (ca şi abaterea medie) sugerează gradul în care valorile cazurilor dintr-o distribuție tind să varieze față de media lor. Valorile mediei şi ale cazurilor sunt folosite în calcul dar *nu*

afectează ordinul de mărime al abaterii standard. Într-un raport de cercetare, media se prezintă de obicei imediat înaintea abaterii standard, astfel încât cititorul să aibă o mai bună imagine asupra a cum arată valorile reale. Adăugând o cantitate fixă la toate valorile unei serii de date, aceasta va afecta media prin creşterea sumei tuturor valorilor, dar nu va afecta abaterea standard pentru acea distribuție. Acest fenomen se întâmplă, de exemplu, când o creştere de 50$ este dată tuturor angajaților unei agenții sociale. Media salariului va fi cu 50$ mai mare, dar cantitatea variației de salariu între angajați va rămâne exact la fel.

Abaterea standard poate fi calculată construind un tabel, în următoarele etape:

1. Se înşiră valorile unei distribuții în coloana *a*
2. Se obține media valorilor din coloana *a*.
3. Se trece valoarea mediei în coloana *b*.
4. Din fiecare valoare din coloana *a* se scade media şi se pune valoarea rezultată în coloana *c*.
5. Se ridică la pătrat fiecare valoare din coloana *c* şi se aşează în coloana *d*.
6. Se adună valorile din coloana *d*.
7. Se împarte suma valorilor din coloana *d* la numărul valorilor din seria de date *a*.
8. Se obține rădăcina pătrată a numărului de la pasul 7. Acest număr reprezintă abaterea standard a valorilor din coloana *a*.

Să calculăm abaterea standard pentru cele 6 valori ale distribuției K.

Distribuția K: 5 6 6 6 7 8

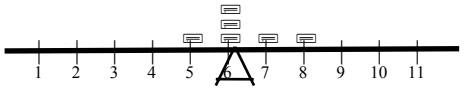
Utilizând cele 8 etape pentru calcularea abaterii standard pentru distribuția dată, putem determina abaterea standard a Distribuției K (Tabelul 5.4).

Tabelul 5.4 Determinarea abaterii standard pentru distribuția K

Pas 1	Pas 3		Pas 4	Pas 5
(a)	(b)		(c)	(d)
Valoare	- Media	=	Abatere față de medie	Pătratele abaterilor față de medie
5	- 6.3	=	-1.3	1.69
6	- 6.3	=	-0.3	0.09
6	- 6.3	=	-0.3	0.09
6	- 6.3	=	-0.3	0.09
7	- 6.3	=	0.7	0.49
8	- 6.3	=	1.7	2.89
				Pas 6 5.34
				Pas 7 5.34 / 6 = 0.89
				Pas 8 $\sqrt{}$ 0.89 = **0.94**

Reprezentarea celor şase valori ale Distribuției K poate fi vizualizată sub forma greutăților pe o bară gradată, aşa cum este reprezentată în figura 5.5.

Figura 5.5 Distribuția de frecvențe K



Să luăm o distribuție care este mai variată decât distribuția K: Distribuția L, tot cu şase valori.

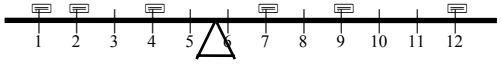
Distribuția L: 1 2 4 7 9 12

Tabelul 5.5 Distribuția de frecvențe L

Pas 1 (a) Valoare	-	Pas 3 (b) Media =	Pas 4 (c) Abatere față de medie	Pas 5 (d) Pătratele abaterilor față de medie
1	-	5.8 =	-4.8	23.04
2	-	5.8 =	-3.8	14.44
4	-	5.8 =	-1.8	3.24
7	-	5.8 =	1.2	1.44
9	-	5.8 =	3.2	10.24
12	-	5.8 =	6.2	38.44
				Pas 6 90.84
				Pas 7 90.84 / 6 = 15.14
				Pas 8 $\sqrt{15.14}$ = **3.89**

Utilizând cele 8 etape putem calcula abaterea standard, pentru Distribuția L (tabelul 5.5). Reprezentarea celor şase valori ale Distribuției L cu ajutorul greutăților pe o bară gradată este prezentată în Figura 5.6 . Din compararea Figurilor 5.5 şi 5.6 reiese evident că distribuția L reflectă mai multă variație decât reflectă Distribuția K. Acest lucru este evidențiat şi de abaterile lor standard.

Figura 5.6 Distribuția de frecvențe L



Calcularea abaterii standard este un proces puțin mai lung. Din fericire nu este necesar să o calculăm manual, deoarece există programe de calculator care o pot face. Cu toate acestea, este important să înțelegem ce înseamnă abaterea standard. O privire aruncată asupra valorilor unor distribuții pe un grafic de tipul figurilor 5.1 - 5.6 ne poate fi de mare ajutor. Dacă nu aveți încă o înțelegere clară asupra conceptului de abatere standard Modulul 6 furnizează clarificări suplimentare.

Varianța şi abaterea standard pot fi calculate manual pornind de la informațiile furnizate de o tabelă de frecvențe. În această situație algoritmul în 8 pași prezentat mai sus se îmbogățește cu un pas suplimentar 5' în care pătratele abaterilor față de medie se înmulțesc cu frecvențele corespunzătoare.

Varianța are proprietatea că este cea mai mică abatere medie pătratică față de o valoare oarecare. Abaterea medie pătratică de la o valoare *a* se calculează, aşa cum arată şi denumirea, prin ridicarea la pătrat a diferențelor dintre valorile individuale şi o mărime constantă, arbitrar aleasă, *a*, urmată de însumarea pătratelor respective şi de raportarea la numărul cazurilor observate.

În plus, abaterea medie pătratică de la o valoare *a,* este mai mare decât varianța exact cu pătratul diferenței dintre cele două valori de referință, adică valoarea mediei minus valoarea *a*. Această proprietate este deosebit de utilă în calculul manual al variațelor în cazul în care valoarea mediei are zecimale, iar ridicările la pătrat devin astfel laborioase. În loc de a calcula diferențele față de valoarea medie, le vom calcula față de o valoare apropiată, convenabil aleasă, apoi vom calcula abaterea medie pătratică față de acea valoare, urmând ca la sfârşit să aducem "corecția" necesară prin scăderea pătratului diferenței dintre valoarea mediei şi valoarea aleasă. Abaterea medie pătratică, astfel "corectată" va fi chiar valoarea variației căutate.

5.6. Utilizarea variației ca măsură a reducerii gradului de nedeterminare. Indicatorul statistic *eta*

Dacă populația statistică este concepută ca fiind formată dintr-o mulțime de grupuri (subpopulații), atunci media generală este egală cu media ponderată a mediilor la nivelul fiecărui grup, ponderile fiind efectivele grupurilor respective. În cazul variației, calculul valorii generale cu ajutorul variațelor obținute la nivelul grupurilor este ceva mai complicat.

Să presupunem că populația este divizată în două subgrupuri, astfel încât avem:

- mediile în cadrul fiecărui grup: m_1, m_2
- varianţele în cadrul fiecărui grup: v_1, v_2
- efectivele grupurilor: n_1, n_2

De asemenea mai avem valorile aceloraşi indicatori statistici la nivelul întregii populaţii:

- media generală: m
- varianţa totală: v
- volumul populaţiei: n

În cazul mediei generale ea este egală cu media mediilor grupurilor ponderate cu efectivele celor două grupuri:

$$m = (m_1 \text{ x } n_1 \ + \ m_2 \text{ x } n_2) / n$$

În cazul varianţei relaţia nu este similară. De regulă, varianţa totală nu este egală cu media ponderată a varianţelor celor două grupe.

$$v_{\ intragrupală} = \ (v_1 \text{ x } n_1 \ + \ v_2 \text{ x } n_2) / n$$

Media ponderată a varianţelor grupurilor (aşa cum s-a calculat în formula de mai sus) poartă denumirea de *varianţă intragrupală* şi este în general mai mică ca valoare decât varianţa totală.

Valorile mediilor grupurilor (m_1 şi m_2) luate cu frecvenţele corespunzătoare (n_1 şi n_2) formează ele însele o variabilă. Această variabilă are şi ea o dispersie ce reflectă împrăştierea mediilor grupale în jurul mediei lor care este, cum am văzut, tocmai media generală a populaţiei. Vom avea deci o varianţă $v_{intergrupală}$ numită *varianţă intergrupală*.

$$v_{\ intergrupală} = \ (\ (m_1 - m)^2 \text{ x } n_1 + \ (m_2 - m)^2 \text{ x } n_2 \ ) \ / \ \ n$$

Există o relaţie simplă între aceste varianţe, după cum ne asigură *teorema de descompunere a varianţei* şi anume *varianţa totală se descompune în suma varianţei intragrupale şi a celei intergrupale*. Utilizând notaţiile de mai sus se poate deci scrie:

$$v \ = v_{\ intragrupală} \ + \ v_{\ intergrupală}$$

Pentru a ilustra această formulă şi a sugera ceva din importanţa ei, pentru modelele explicative, vom trata un exemplu. Să prepunem că o într-o comunitate formată din 60 de familii, 15 beneficiază de ajutor social. Celelalte 45 nu sunt incluse într-un astfel de program. Caracteristica "a fi sau nu beneficiar de ajutor social" împarte deci populaţia de familii în două grupuri. Ea este *variabila ce serveşte pentru clasificare*. Ajutorul social se acordă pe baza unui set de criterii complexe, dintre care unul ar putea fi numărul de membrii în familie. Să mai presupunem că, la rândul lor, aceste familii se distribuie ca în tabelele 5.6 şi 5.7. Suma frecvenţelor corespunzătoare va reproduce, desigur, distribuţia după numărul de membrii a familiilor întregii comunităţi, tabelul 5.8.

Tabelul 5.6. Distribuţia familiilor cu ajutor social după numărul de membri

Număr de membrii	Număr de familii
1	1
3	1
4	3
5	4
6	4
7	1
8	1
Total	15

Tabelul 5.7. Distribuția familiilor fără ajutor social după numărul de membri

Număr de membrii	Număr de familii
1	4
2	10
3	22
4	9
Total	45

Tabelul 5.8. Distribuția familiilor după numărul de membri

Număr de membrii	Număr de familii
1	5
2	10
3	23
4	12
5	4
6	4
7	1
8	1
Total	60

Făcând calculul mediei şi al varianței pentru fiecare din cele două grupuri (tabele) şi pentru întreaga comunitate, obținem:

- pentru grupul familiilor ce beneficiază de ajutor social:

  m_1 = 5 membri ; v_1 = 2,66

- pentru grupul format din familiile care nu beneficiază de ajutor social:

  m_2 = 2,8 membri ; v_2 = 0,74

- pentru întreaga comunitate:

  m = 3,35 membri ; v = 2,13

Se vede clar că cele două categorii de familii diferă net atât prin numărul de membri cât şi prin gradul de dispersie. Dar nu acest aspect ne interesează aici, ci faptul că intrând cu analiza în interiorul celor două subpopulații, vom obține o dispersie medie (adică o nedeterminare a situației) a acestora, care este mai mică decât cea calculată pe ansamblul comunității. Astfel, varianța intragrupală va fi:

$$v_{intragrupală} = (15v_1 + 45\ v_2) / 50 = (15 \times 2{,}66 + 45 \times 0{,}74) / 60 = 1{,}23$$

Celălalt membru al sumei, varianța intergrupală, este indicatorul de dispersie al variabilei care ia ca valori cele două medii ale grupurilor, iar ca frecvențe, efectivele acestora. Prin urmare, vom avea:

$$\begin{aligned} v_{intergrupală} &= (\ 15(m_1-m)^2 + 45(m_2-m)^2\ ) / 60 \\ &= (\ 15(5-3{,}35)^2\ ) + 45(2{,}8-3{,}35)^2\ ) / 60 \\ &= 0{,}90 \end{aligned}$$

Deci, din variația totală de 2,13, aproximativ 58% (1,23 / 2,13 x 100 = 58) este varianță intragrupală, reflectând gradul de nedeterminare ce rămâne, în medie, la nivel de grupuri, pe când circa 42% (adică 0,90 / 2,13 x 100 = 42) este pusă pe seama variației mediilor celor două grupuri alese şi vom

spune că ea se consideră a fi explicată prin analiza la nivel de grupuri, grupuri care pot fi, evident, interpretate drept clasele unei alte variabile, cu funcție predictivă sau explicativă pentru variabila studiată.

Raportul dintre varianța intergrupală şi varianța totală (exprimată în procente) se numeşte eta şi este un indicator statistic care arată cu cât s-a redus gradul de nedeterminare a unei variabile (de explicat) prin introducerea în analiză a unei alte variabile (explicative).

În exemplul nostru, prin introducerea variabilei de clasificare "a beneficia sau nu de ajutor social" în studiul variabilei "număr de membri în familie" cantitatea totală de nedeterminare 2,13 s-a redus la 1,23. Gradul de nedeterminare s-a redus cu 0,90 (varianța intergrupală), ceea ce reprezintă aproximativ 42% (*eta* = 0,90 /2,13 * 100 = 42,25) din varianța totală. Putem afirma, deci, că mărimea familiei a fost un criteriu important în acordarea ajutorului social, ea a contat în proporție de 42%. Restul criteriilor (venituri, vârsta, proprietăți etc.) explică celelalte 58% din variația totală.

Observație. Şi din compararea directă a mediilor putem vedea că există diferențe importante între mărimile medii ale celor două grupuri de familii (5 membrii/familie pentru grupul celor cu ajutor social față de numai 2,8 pentru grupul celor fără ajutor social). Abaterea standard în cadrul grupurilor arată că şi din punctul de vedere al omogenității / eterogenității cele două grupuri de familii diferă. Cele cu ajutor social sunt mult mai diversificate (abatere standard 1,63) față de celelalte (abatere standard 0,86).

REZUMAT

În acest capitol, am discutat procesele rezumării şi descrierii informației. Împrăştierea în raport cu tendința centrală clarifică imaginea unui set de date. Indicatorii împrăştierii, asemenea amplitudinii sau a abaterii intercuartile sunt uşor de calculat, dar pot fi înțelese eronat. Formulele după care se calculează nu iau în considerare toate valorile colectate. Dacă există nişte valori deviante sau dacă informațiile sunt de nivel interval sau rapoarte se preferă abaterea standard. Cititorii unui raport de cercetare în care s-a furnizat media şi abaterea standard a unei variabile vor fi capabili să reconstruiască o imagine completă a datelor. Aşa cum am explicat în modulul despre indicatorii tendinței centrale, dacă criteriile pentru utilizarea mediei nu sunt întrunite, trebuie să considerăm că este necesară folosirea unui alt indicator. În mod similar dacă nu se regăsesc toate criteriile pentru folosirea abaterii standard, va trebui să folosim un indicator al împrăştierii mai puțin senzitiv, ca de exemplu abaterea intercuartilă sau amplitudinea.

Întrebări pentru studiu

1. Cum ne ajută indicatorii variabilității să completăm descrierea datelor noastre?
2. De ce este amplitudinea un indicator nestabil al variabilității?
3. Ce face ca abaterea intercuartilă să fie un indicator al variabilității mai stabil decât amplitudinea, pentru date de nivel interval sau rapoarte?
4. Care indicatorii variabilității iau în considerare toate valorile variabilei pentru calcularea lor?
5. Cum poate o valoare deviantă (foarte înaltă sau foarte scăzută) să distorsioneze abaterea medie pentru un grup de date.
6. Cum este afectată abaterea standard prin adăugarea numărului 10 la fiecare din valorile unei variabilele?
7. Cum se poate compara un poligon de frecvențe a unei serii de date cu media 10 şi abaterea standard 3, cu un poligon de frecvențe a altei serii de date tot cu media 10 şi o abatere standard de 12, dacă ele sunt reprezentate pe acelaşi sistem de axe rectangulare?
8. Dacă câteva valori foarte deviante dintr-un set de date măsurate pe o scală de nivel interval distorsionează rău abaterea standard, care indicator al variabilității trebuie folosit?
9. Ştim că abaterea intercuartilă este adesea folosită atunci când mediana este cel mai potrivit indicator al tendinței centrale. Ce indicator al variabilității este folosit atunci când media este cel mai potrivit indicator de poziție?
10. Un asistent social dintr-un centru comunitar de sănătate mentală se întreabă dacă terapia emoțională este mai adecvată pentru femei decât pentru bărbați. El întreprinde un studiu bazat pe cazurile a 30 de clienți tratați (20 femei şi 10 bărbați) şi înregistrează progresele obținute în sănătatea mentală a pacienților pe o scală de la 0 la 4 puncte. În urma calculelor au rezultat următoarele:
    - Toți clienții împreună: media 2,53 puncte; varianța 1,85; abaterea standard 1,36
    - Femeile: media 3,30 puncte; varianța 0,71; abaterea standard 0,84

- Bărbații: media 1,00 punct; varianța 0,60; abaterea standard 0,77

Comentați rezultatele obținute și răspundeți la întrebarea asistentului social.

Exercițiu pentru acasă 5.

Paul lucrează la serviciul public de asistență socială din cadrul primăriei, având în evidență un număr de 50 de familii solicitante de ajutor social. Unul din criteriile de acordare a ajutorului social se referă la starea generală de sănătate a membrilor familiei. Iată cum arată distribuția familiilor de care se ocupă Paul, în funcție de un punctaj pe o scală de evaluare a stării de sănătate, cu maxim 5 (probleme grave de sănătate) și minimum 1 (stare de sănătate bună):

Tabel 1. Distribuția familiilor după starea de sănătate

Nivel de sănătate	1	2	3	4	5	Total
Nr. familii	3	5	20	10	12	50

Paul vrea să verifice dacă starea generală de sănătate a membrilor familiei ar putea constitui un criteriu important în acordarea ajutorului social. În acest sens, Paul consultă anchetele sociale ale celor 50 de familii și le regrupează după acest criteriu. Cele două subgrupuri care rezultă se distribuie astfel:

Tabel 2. Distribuția familiilor care beneficiază de ajutor social

Nivel sănătate	1	2	3	4	5	Total
Nr. familii	3	1	12	4	10	30

Tabel 3. Distribuția familiilor care nu beneficiază de ajutor social

Nivel sănătate	1	2	3	4	5	Total
Nr. familii	-	4	8	6	2	20

Calculati mediile si varianțele pentru a afla la ce concluzie a ajuns Paul. Comentați pe scurt rezultatul obținut.

Modul 6

Distribuția normală

Capitolele precedente au prezentat indicatorii principali ai distribuțiilor de frecvențe privind tendința centrală şi variabilitatea. Există o formă specială a distribuțiilor de frecvențe cunoscută sub numele de distribuție normală. Cunoaşterea distribuției normale este esențială pentru înțelegerea inferenței statistice şi pentru testarea ipotezelor.

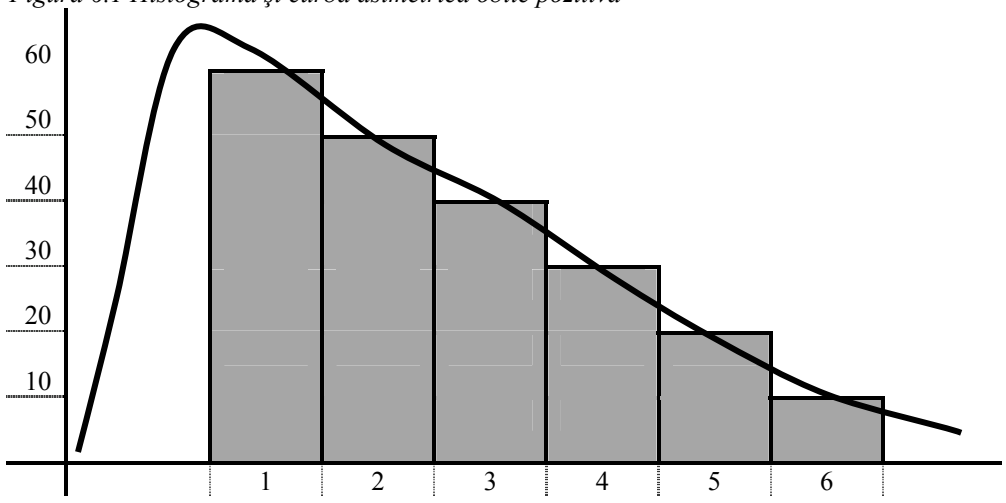
Să considerăm o distribuție de frecvențe ca în Tabelul 6.1. Aşa cum am văzut în discuția asupra poligoanelor de frecvențe din Modulul 3, distribuțiile de frecvențe pot fi prezentate grafic. Astfel de grafice pot lua diferite forme. De exemplu, unele pot arăta cum un număr mare de cazuri se concentrează în jurul unor valori mai mici şi doar câteva se distribuie între valorile mai mari. Astfel este şi distribuția prezentată în Tabelul 6.1.

Distribuția frecvențelor din Tabelul 6.1 poate fi uşor reprezentată grafic într-o histogramă, aşa ca în Figura 6.1. Linia continuă care uneşte mijloacele segmentelor din vârful intervalelor din histogramă se numeşte *curba frecvențelor*.

Tabelul 6.1 Distribuție de frecvențe cumulate

Rezultate	Frecvențe absolute	Frecvențe cumulate (Mici-Mari)	Frecvențe cumulate (Mari-Mici)
1	60	60	210
2	50	110	150
3	40	150	100
4	30	180	60
5	20	200	30
6	10	210	10

Figura 6.1 Histogramă şi curbă asimetrică oblic pozitivă



Distribuțiile ca cele din Tabelul 6.1 şi Figura 6.1 sunt numite *oblice* sau *asimetrice*. *Oblicitatea* (skewness) arată că distribuția nu este simetrică - adică valorile marginale nu sunt uniform repartizate în ambele capete. Curba din Figura 6.1 este oblică la dreapta deoarece frecvențele se răresc mai mult la dreapta. O curba care este oblică la dreapta este numită *oblic pozitivă* sau *pozitiv asimetrică*.

Dacă am completa invers frecvenţele, din Tabelul 6.1, şi le-am aranja ca în Tabelul 6.2, am avea o distribuţie de frecvenţe ca cea reprezentată de curba din Figura 6.2. Această distribuţie este tot oblică. De această dată, frecvenţele se răresc la stânga. O curbă care este oblică la stânga este denumită *oblic negativă* sau *negativ asimetrică*.

Tabelul 6.2 Distribuţie de frecvenţe cumulate

Rezultate	Frecvenţe absolute	Frecvenţe cumulate
1	10	10
2	20	30
3	30	60
4	40	100
5	50	150
6	60	210

Figura 6.2 Histogramă şi curbă asimetrică oblic negativă

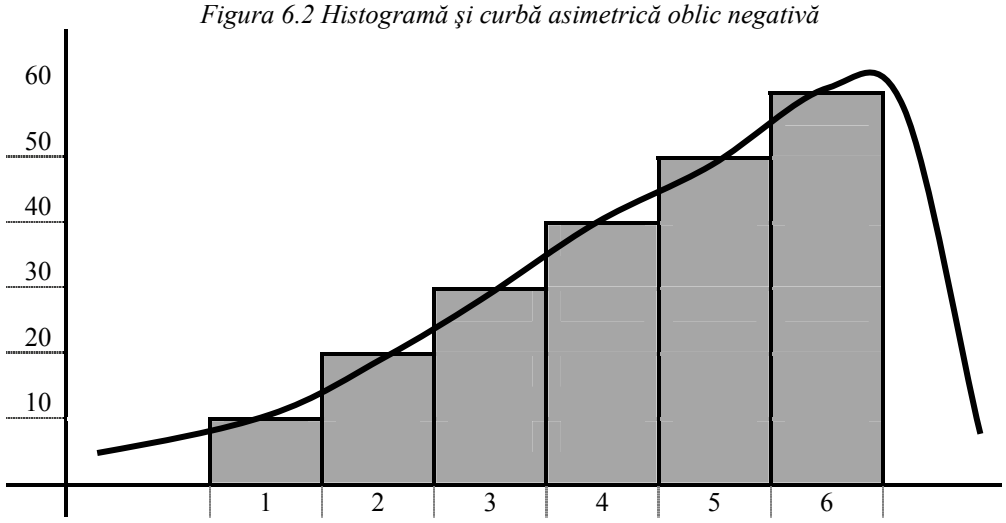
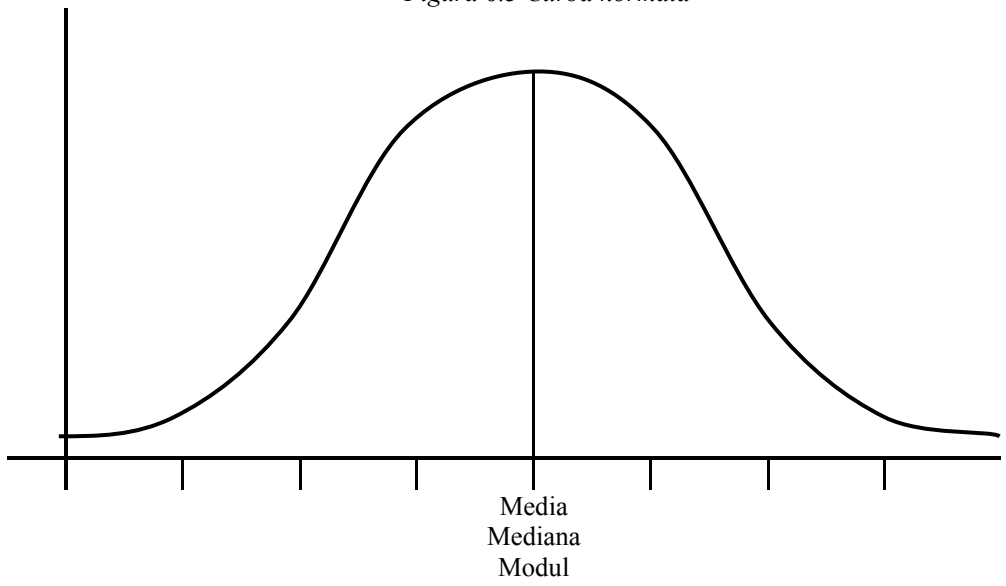


Figura 6.3 Curba normală



Când o curbă are aceiaşi oblicitate şi la stânga şi la dreapta, la fel ca cea din Figura 6.3, se spune că este *simetrică*, şi că distribuţia de frecvenţe pe care o reprezintă este simetrică. În plus curba din Figura 6.3 este descrisă ca fiind *normală.* Nu toate curbele simetrice sunt normale, dar toate curbele normale sunt simetrice. Cuvântul normal se referă la un tip specific de curbă şi o distribuţie cu proprietăţi matematice bine precizate. Alt termen pentru o curbă normală este *curba sub formă de clopot*. Această formă, este deosebită de alte curbe de frecvenţe simetrice şi arată ca cea din figura 6.3.

6.1. Proprietăţi ale distribuţiei normale

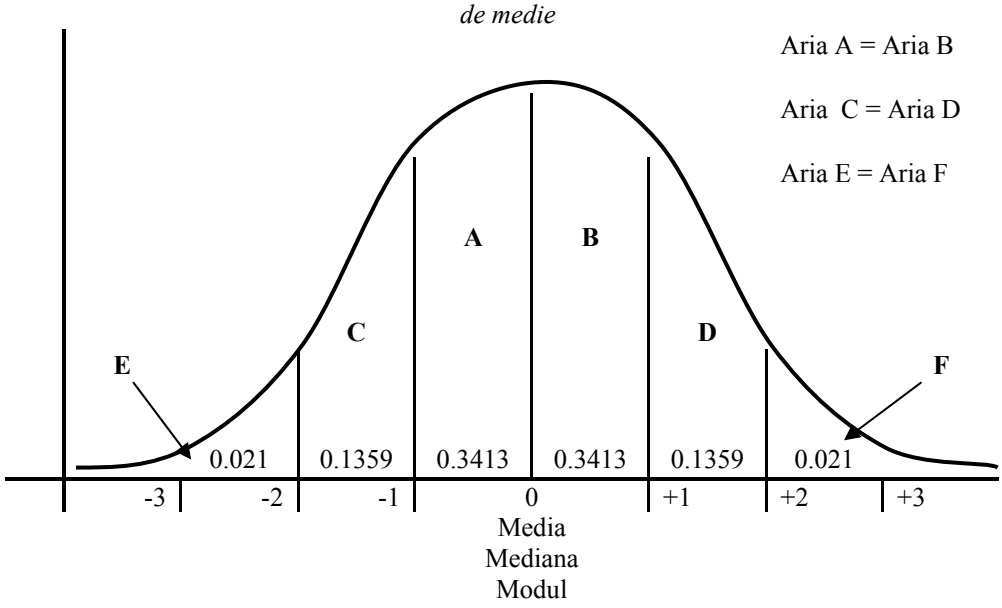
Distribuţia normală este o distribuţie teoretică, matematică, şi ea nu este una reală, empirică. Ea este un fel de „etalon" la care se raportează distribuţiile reale. Indiferent de mărimea tendinţei centrale şi a dispersiei toate distribuţiile normale au aceleaşi proprietăţi.

1. Media, mediana şi modul (trei indicatori ai tendinţei centrale) distribuţiei normale, *se găsesc toate în acelaşi punct* - adică în punctul de pe axa orizontală corespunzător cel mai înalt punct al curbei din centrul distribuţiei. Două curbe normale pot avea medii şi abateri standard diferite. La curbele oblice (asimetrice), media, mediana şi modul se găsesc în puncte diferite .

2. Distribuţia normală *conţine şase abateri standard* (trei de fiecare parte a mediei). Figurile 6.3 şi 6.4 reprezintă distribuţii normale şi ilustrează aceste proprietăţi. Aşa cum am văzut, ca să comparăm două distribuţii diferite avem nevoie să luăm în considerare atât mediile cât şi abaterile lor standard. Două distribuţii normale pot avea medii identice dar abateri standard diferite, pot avea medii diferite dar abateri standard identice sau pot diferi atât prin medii cât şi prin abateri standard.

3. Aşa cum am subliniat, distribuţia normală este simetrică şi sub formă de clopot. *Marginile distribuţiei se extind către infinit* - sfârşitul distribuţiei întotdeauna se apropie de axa orizontală dar niciodată nu o întâlneşte. Din acest motiv, marginile unei curbe normale în toate figurile din acest capitol sunt prezentate ca ne-atingând axa orizontală. Această proprietate reprezintă posibilitatea ca un număr foarte mic de valori se pot găsi dincolo de trei abateri standard faţă de medie. De asemenea reflectă faptul că la un nivel înalt de abstractizare, nu există niciodată un univers total (sau o populaţie totală), deoarece întotdeauna există o posibilitate de schimbare, ca de exemplu apariţia unor noi cazuri sau dispariţia altora mai vechi. Există populaţii care nu sunt niciodată statice; ele întotdeauna evoluează.

Figura 6.4. Suprafeţe delimitate de curba normală şi de un număr întreg de abateri standard distanţă faţă de medie



4. Aria delimitată de curba normală, de axa orizontală şi de perpendicularele ridicate de la un punct de pe axa orizontală şi medie este egală cu aria delimitată de perpendiculara ridicată dintr-un punct aflat la aceeaşi distanţă faţă de medie dar de cealaltă parte a ei, pe axă. De exemplu, o distanţă de o abatere

standard la stânga mediei este egală cu o distanță de o abatere standard la dreapta față de medie, iar suprafețele de sub curbă care corespund acestor distanțe, atât în stânga cât şi în dreapta mediei, sunt egale. Aceasta este proprietatea de *simetrie* a curbei. Figura 6.4 ilustrează acest lucru foarte clar.

6.2. Valori Z şi distribuția normală standardizată

Ori de câte ori întâlnim valori ale unor variabile bazate pe măsurători luate din două populații diferite, putem fi puşi în situația de a nu putea face comparări directe. De exemplu, nu putem compara în mod direct clasa socio-economică a doi indivizi, dacă cunoaştem doar venitul în dolari câştigat în două țări diferite. Compararea este posibilă doar când ştim unde se situează venitul în dolari pentru fiecare caz în parte față de alte cazuri din grupurile lor respective. Acest surplus de informații este necesar dacă mediile şi abaterile standard pentru cele două populații nu sunt identice. Ceea ce este foarte rar întâlnit astăzi în practica cercetării de asistență socială.

Putem depăşi însă dificultățile prin convertirea valorilor luate de la diferite populații la un *standard comun*. Un asemenea standard poate fi găsit cu ajutorul distribuției normale şi a valorilor Z. Compararea valorilor din diferite populații după o astfel de conversie nu mai creează probleme, fiecare populație este redusă la un acelaşi standard şi este măsurată în unități de abateri standard (valori Z).

Valorile Z sunt valorile brute (aşa cum au fost ele măsurate) *transformate în unități de abateri standard*. Fiecare dată brută poate fi dată ca echivalentul unei valori Z care ne spune la câte abateri standard se găseşte față de medie. Pozițiile relative a două valori Z luate de la două distribuții normale diferite pot fi după aceea comparate.

Datorită naturii distribuției normale, valorile Z pot fi transformate într-un termen matematic mult mai familiar, *centila*. O valoare aflată la centila 75% trebuie să fie mai mare decât trei sferturi din totalitatea valorilor din distribuție[1] şi mai mică decât un sfert din valori. Dacă de exemplu, Emil a realizat la un test de statistică un punctaj situat la centila 82%, iar Maria a realizat la un alt test de statistică un punctaj situat la centila 92%, este de presupus că Maria s-a situat mai bine la test decât Emil, chiar dacă amândoi au primit 75 de puncte la respectiva examinare. Pentru a schimba o valoare brută într-o valoare Z, se foloseşte următoarea formulă:

$$\text{valoarea Z} = \frac{\text{valoarea brută - media}}{\text{abaterea standard}}$$

Când cunoaştem media şi abaterea standard a distribuției din care fac parte valorile brute, putem calcula valorile Z. Aşa cum am sugerat mai sus, o valoare Z poate fi percepută ca numărul de abateri standard față de media distribuției. O valoare 1 pentru Z reprezintă o abatere standard de la medie, o valoare 2 reprezintă două abateri standard peste medie şi aşa mai departe. O valoare Z poate fi pozitivă sau negativă în concordanță cu poziția valorii particulare față de medie. O valoare Z pozitivă este o valoare situată peste medie, iar o valoare Z negativă este una sub medie.

Valorile Z a oricărei distribuții normale iau forma *distribuției normale standardizate*, adică o distribuție teoretică, complet cunoscută, cu toți indicatorii statistici cunoscuți. Astfel media este egală cu 0, abaterea standard este egală cu 1, indicatorul de simetrie este 0, la fel ca şi valoarea indicatorului de boltire.

În plus *aria totală delimitată de curbă normală standardizata este egală cu unu* (1,0). Unu (întregul*),* aşa cum folosim termenul aici, se referă la faptul că toate valorile Z (100%) ale unei variabile distribuite normal, vor cădea dedesubtul curbei normale. Părți din acest întreg (câteodată referită prin *unitate*) sunt reprezentate ca proporții din aria totală aflată dedesubtul curbei şi înăuntrul distribuției. De exemplu, 50 de procente ale suprafeței delimitată de curba normală, aflate la stânga mediei, reprezintă acea jumătate a valorilor care cad sub medie. Într-o distribuție normală (ca cea prezentată în Figura 6.3) toate valorile se regăsesc dedesubtul curbei.

Aşa cum se vede şi în Figura 6.4, suprafața delimitată de curba normală şi verticalele Z = +1 şi medie este egală cu suprafață cuprinsă între verticala Z = -1 şi medie şi este egală cu 0,3413 (sau în procente, 34,13%). Aria suprafeței de sub curbă delimitată de verticalele ridicate din oricare două puncte

[1] Acest lucru este valabil numai in cazul în care distribuția reală este „normală", nu şi dacă aceasta are o altă formă.

(valori Z) este egal cu procentul cazurilor ale căror valori Z se găsesc între cele două verticale. De exemplu, 34,13% din totalul valorilor dintr-o distribuție normală cad între medie şi Z = +1; 34,13% între medie şi Z = -1; 13,59% între Z = +2 şi Z = +1, 13,59 procente între Z = -2 şi Z = -1, şi aşa mai departe.

O mare parte a suprafeței de sub curbă şi deci şi un mare număr al cazurilor dintr-o distribuție (68,26%) cad în centru (între valorile Z = -1 şi Z = +1). Doar o mică parte a suprafeței de sub curbă şi un procent mic de cazuri dintr-o distribuție normală (mai puțin de o jumătate de procent) cad în afara valorilor Z = -3 şi Z = +3. Aceasta înseamnă că, într-o distribuție normală, valorile care cad sub sau peste trei abateri standard față de medie sunt foarte rare.

Tabelul 6.3 Aria cuprinsă între curba normală, medie şi valoarea Z

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.00	0.40	0.80	1.20	1.60	1.99	2.39	2.79	3.19	3.59
0.1	3.98	4.38	4.78	5.17	5.57	5.96	6.36	6.75	7.14	7.53
0.2	7.93	8.32	8.71	9.10	9.48	9.87	10.26	10.64	11.03	11.41
0.3	11.79	12.17	12.55	12.93	13.31	13.68	14.06	14.43	14.80	15.17
0.4	15.54	15.91	16.28	16.64	17.00	17.36	17.72	18.08	18.44	18.79
0.5	19.15	19.50	19.85	20.19	20.54	20.48	21.23	21.57	21.90	22.24
0.6	22.57	22.91	23.24	23.57	23.89	24.22	24.54	24.86	25.17	25.49
0.7	25.80	26.11	26.42	26.73	27.04	27.34	27.64	27.94	28.23	28.52
0.8	28.81	29.10	29.39	29.67	29.95	30.23	30.51	30.78	31.06	31.33
0.9	31.59	31.86	32.12	32.38	32.64	32.90	33.15	33.40	33.65	33.89
1.0	34.13	34.38	34.61	34.85	35.08	35.31	35.54	35.77	35.99	36.21
1.1	36.43	36.65	36.86	37.08	37.29	37.49	37.70	37.90	38.10	38.30
1.2	38.49	38.69	38.88	39.07	39.25	39.44	39.62	39.80	39.97	40.15
1.3	40.32	40.49	40.66	40.82	40.99	41.15	41.31	41.47	41.62	41.77
1.4	41.92	42.07	42.22	42.36	42.51	42.65	42.79	42.92	43.06	43.19
1.5	43.32	43.45	43.57	43.70	43.83	43.94	44.06	44.18	44.29	44.41
1.6	44.52	44.63	44.74	44.84	44.95	45.09	45.15	45.25	45.35	45.45
1.7	45.54	45.64	45.73	45.82	45.91	45.99	46.08	46.16	46.25	46.33
1.8	46.41	46.49	46.56	46.64	46.71	46.78	46.86	46.93	46.99	47.06
1.9	47.13	47.19	47.26	47.32	47.38	47.44	47.50	47.56	47.61	47.67
2.0	47.72	47.78	47.83	47.88	47.93	47.98	48.03	48.08	48.12	48.17
2.1	48.21	48.26	48.30	48.34	48.38	48.42	48.46	48.50	48.54	48.57
2.2	48.61	48.64	48.68	48.71	48.75	48.78	48.81	48.84	48.87	48.90
2.3	48.93	48.96	48.98	49.01	49.04	49.06	49.09	49.11	49.13	49.16
2.4	49.18	49.20	49.22	49.25	49.27	49.29	49.31	49.32	49.34	49.36
2.5	49.38	49.41	49.41	49.43	49.45	49.46	49.48	49.49	49.51	49.52
2.6	49.53	49.55	49.56	49.57	49.59	49.60	49.61	49.62	49.63	49.64
2.7	49.65	49.66	49.67	49.68	49.69	49.70	49.71	49.72	49.73	49.74
2.8	49.74	49.75	49.76	49.77	49.77	49.78	49.79	49.79	49.80	49.81
2.9	49.81	49.82	49.82	49.83	49.84	49.84	49.85	49.85	49.86	49.86
3.0	49.87									
3.5	49.85									
4.0	49.997									
5.0	49.99997									

Dacă o valoare Z, corespunzătoare unei valori brute date, nu este un număr întreg (Z = -3, Z = +2, etc.) datele înscrise în suprafețele distribuției normale din Figura 6.4 nu sunt suficiente pentru a transforma o valoare Z într-o centilă. Din fericire există tabele care ne ajută în transformarea unei valori Z fracționare (Z=0,28, Z= 2,15, etc.), în centile.

Tabelul 6.3 cuprinde valori Z întregi şi fracționare. El ne furnizează aria de sub curba normală cuprinsă între valoarea Z (întreagă sau fracționară) şi medie, pentru orice distribuție normală. Cifra care

reprezintă partea întreagă şi prima cifră zecimală a valorii Z sunt etichetele liniilor din tabel, şi se citesc în partea stângă din prima coloană a tabelului 6.3, în timp ce cea de a doua cifră zecimală a numărului fracţionar se citeşte din capul de tabel. Numărul găsit în tabel la intersecţia liniei şi coloanei corespunzătoare unei valori Z, este proporţia din suprafaţa curbei normale cuprinsă între medie şi valoarea Z. Exemplele din Tabelul 6.4 arată cum poate fi utilizat Tabelul 6.3.

Tabelul 6.4 Exemple de valori Z şi ariile lor corespunzătoare

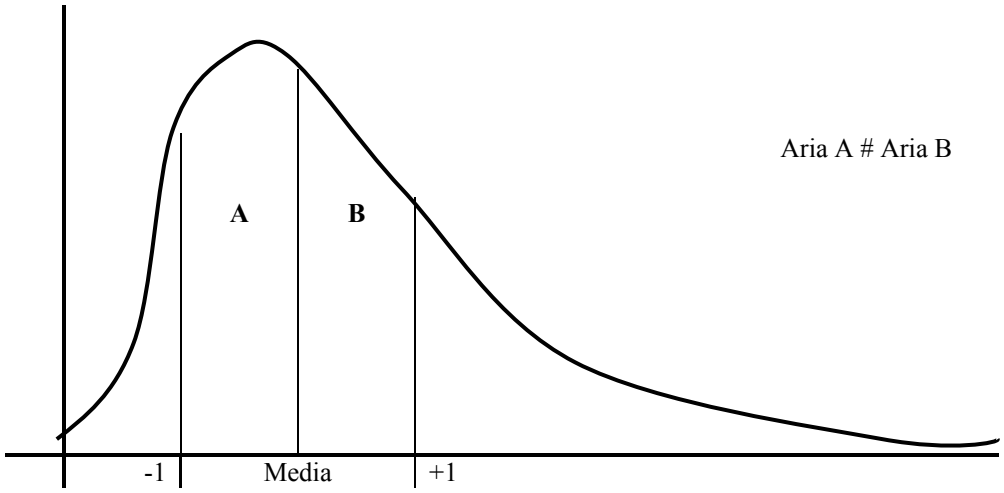
Valoare Z	Linia	Coloana	Aria corespunzătoare
.12	0.1	.02	4.78
1.78	1.7	.08	46.25
-2.90	2.9	.00	49.81
1.15	1.1	.05	37.49
-1.15	1.1	.05	37.49

Semnul (+ sau -) indică doar dacă valoarea Z este sub sau peste medie. Aria proporţională indusă de valorile Z de aceeaşi mărime, dar semne diferite, este aceeaşi. Pentru valori brute peste medie (având o valoare Z pozitivă), *adunăm* aria cuprinsă între medie şi valoarea Z la 50,00 pentru a afla la ce centilă se găseşte acea valoare (brută). Pentru valori brute sub medie, *scădem* din 50,00 aria dintre medie şi valoarea Z pentru a afla la ce centilă se găseşte acea valoare. De exemplu (vezi Tabelul 6.4), o valoare brută transformată într-o valoare Z de 0,12 va cădea aproape de centila 55% (50,00 + 4,78 = 54,78). Aşa cum arată Tabelul 6.4, o valoare Z de 1,78 va cădea aproape de centila 96% (50,00 + 46,25 = 96,25); o valoare Z de -1,15 va cădea aproape de centila 13% (50,00 - 37,45 = 12,51) şi aşa mai departe.

Avantajele şi dezavantajele valorilor Z

Primul avantaj al valorilor Z este acela că ele pot fi folosite pentru a compara valori brute luate de la diferite distribuţii. Datele trebuie să fie cel puţin de nivel interval.

Figura 6.5. Compararea suprafeţelor corespunzătoare unei abateri standard la curbele asimetrice



Cel mai mare dezavantaj al valorilor Z este că presupun o distribuţie normală. Când această calitate nu este întâlnită, o valoare Z nu poate fi interpretată în termenii unei proporţii standardizate a distribuţiei pentru care a fost calculată. Când distribuţia este oblică, aria cuprinsă într-o abatere standard la stânga faţă de medie nu este egală cu aria aflată la aceiaşi distanţă dar la dreapta mediei. De exemplu, distribuţia din Figura 6.5 este oblic pozitivă. Aria A nu este egală cu aria B, chiar dacă fiecare arie corespunde unei abateri standard faţă de medie.

Utilizarea distribuției normale în teste standardizate

Distribuția normală se utilizează în testele standardizate, asemenea testelor de inteligență IQ (Intelligence Quotient), iar în SUA chiar şi pentru testele de aptitudini şcolare SAT (Scholastic Aptitude Tests) sau examenul de bacalaureat GRE (Graduate Record Exams). Dealungul anilor aceste teste au fost îmbunătățite şi adaptate frecvent astfel ca valorile rezultate, prin testarea unui mare număr de persoane, să tindă să corespundă unui model cu o medie şi o abatere standard consistente. Astfel de rezultatele vor forma întotdeauna o distribuție normală. Testele IQ în general sunt structurate să aibă o medie de 100 puncte şi o abatere standard de 15 sau 16 puncte, depinzând de test. Dacă înțelegem principiile şi caracteristicile legate de distribuția normală, cu aceste informații putem să transformăm data brută IQ într-o valoare Z şi apoi în centile, folosind Tabelul 6.3. De exemplu, un coeficient de inteligență cu o valoare Z de +1 (115 sau 116 puncte) va cădea aproape de centila 84%. Transformarea inversă, din centile în valori brute, este de asemenea posibilă.

De exemplu, un individ aflat printre primii 5% dintre cei mai inteligenți oameni realizează un scor IQ de cel puțin 126,25 (100 + 1,64 * 16) puncte. Rezultatul se obține aplicând formula de calcul al valorii brute, derivată din formula valorilor Z:

valoarea brută = media + abaterea standard * valoarea Z

unde media este 100, abaterea standard 16, iar valoarea Z de 1,64 am aflat-o din Tabelul 6.3 în dreptul ariei 45 (de fapt aria cea mai apropiată de 45 în tabel este de 45,09).

Valorile SAT şi GRE au fost astfel proiectate ca ele să combine valori verbale şi matematice pentru un număr mare de elevi, astfel încât să formeze o curbă normală cu media 1000 şi abaterea standard 200. Deoarece în mod virtual toate valorile vor cădea practic între +3 şi -3 abateri standard de la medie, valoarea cea mai scăzută va fi cu de trei ori 200, adică cu 600 puncte, sub medie (adică 400). Cea mai mare valoare (aflată la centila 100%) va cădea la 1600 puncte. În realitate (în S.U.A), datorită scăderii valorilor SAT şi GRE înregistrate în ultimii ani, media a scăzut considerabil sub 1000 pentru un an dat, dar valorile continuă să formeze o distribuție care este aproape simetrică şi încă acceptată ca normală.

6.3. Utilizarea valorii Z în comparații: un exemplu practic

Valorile Z sunt folosite în special pentru compararea datelor brute a două cazuri când s-au utilizat scale de măsură diferite pentru aceeaşi variabilă. Ele ne permit să comparăm poziția relativă a unui caz (centila măsurii sale) cu poziția relativă a celuilalt caz (centila măsurii sale). Presupunând că două grupuri sunt normal distribuite după o variabilă, vom putea face o comparare validă a două cazuri. Un exemplu ne va ajuta să ilustrăm această utilizare.

Dorina, un asistent social dintr-o agenție de consiliere familială, a condus un grup de studenți de colegiu diagnosticați că au trăit experiențe de anxietate acută. În trecut, membrii grupului erau selectați pe baza rezultatelor la testul Scalei de Anxietate A, test aplicat tuturor studenților luați în considerare. Această scală are o medie de 70 şi o abatere standard de 10. Numai clienții care realizau peste 80 puncte pe Scala de Anxietate A aveau permisiunea să se alăture grupului.

S-a întâmplat la un moment dat să apară un loc vacant în grup. Dorina a căutat printre dosarele cazurilor active din agenția sa şi a observat că cea mai înaltă valoare realizată de potențialii membrii ai grupului este de 78 (Gina). Ea a primit un raport de la o altă agenție familială despre un nou client (Toma) recent mutat în oraşul său şi care necesita asistență. Scrisoarea de transfer arăta că Toma, care suferea de anxietate, a realizat o valoare de 66 puncte pe Scala de Anxietate B - o scală care era diferită de cea pe care o folosise Dorina. Scrisoarea mai indică faptul ca Scala de Anxietate B avea o medie de 50 şi o abatere standard de 12.

Ambele scale de anxietate erau teste standardizate. Bazându-se pe cunoştințele sale asupra distribuției normale şi pe informațiile primite în scrisoarea de recomandare, Dorina s-a văzut nevoită să îl re-testeze pe Toma. S-a decis să folosească valoarea Z pentru a determina care dintre studenții Gina sau Toma, este un candidat mai potrivit pentru locul vacant din grup. Pentru a simplifica decizia sa, Dorina a construit un tabel comparativ cu datele sale (Tabelul 6.5). Apoi a calculat valoarea Z, pentru ambii potențiali clienți, ceea ce i-a permis să determine centilele pentru fiecare valoare.

Tabelul 6.5 Date comparative: Indicatorii celor două scale şi valorile realizate de clienţi

Valori	Scala de Anxietate A (Gina)	Scala de Anxietate B (Toma)
Valoarea brută	78	66
Media scalei	70	50
Abaterea standard	10	12

Valoarea Z (Gina) = (Valoarea realizată - Media) / Abaterea standard
= (78 - 70) / 10 = 8 / 10 = 0,8

Valorii Z = .80 îi corespunde 28.81 în Tabelul 6.3

28.81 (aria dintre valoare şi medie)
\+ 50.00 (aria din stânga mediei)
78.81 = centila 79 % (Scala A)

Valoarea Z (Toma) = (Valoarea realizată - Media) / Abaterea standard
= (66 - 50) / 12 = 16 / 12 = 1,33

Valorii Z = 1.33 îi corespunde 40.82 în Tabelul 6.3

40.82 (aria dintre valoare şi medie)
\+ 50.00 (aria din stânga mediei)
90.82 = centila 91 % (Scala B)

Bazându-se pe analiza comparativă a valorilor Z, Dorina îl alege pe Toma pentru grup. Nivelul înalt al anxietăţii (pe baza scalei sale) l-a făcut pe Toma un candidat evident. Mai mult, ea nu a mai fost nevoită să slăbească criteriile de admisibilitate pentru a o include pe Gina în grup (care se situa la centila 79%) când se cerea o valoare de cel puţin 80 puncte (centila 84%) sau mai mult (pe baza pe scalei ei). De notat că valoarea de 80 pe Scala A este comparabilă cu o valoare de 62 pe Scala B, deoarece ambele cad la punctul Z = +1 (centila 84%). Valoarea Z corespunzătoare punctajului realizat de Toma a fost deasupra acestui punct şi bineînţeles deasupra celui realizat de Gina .

REZUMAT

În acest capitol am prezentat un tip special de analiză a datelor, care poate fi utilizat când valorile unei variabile formează o distribuţie normală. Prin transformarea unei valori brute, a unei variabile normal distribuite, în valoare Z, putem compara valorile individuale din două seturi diferite de valori sau din două subseturi, prin observarea modului în care fiecare valoare se raportează la celelalte valori din grupul propriu. De asemenea înţelegând distribuţia normală, putem să vizualiză locul în care o valoare dată cade faţă de celelalte dintr-o populaţie de volum mare. Prin folosirea centilelor, putem determina procentul de valori sub care, sau peste care, se situează o anumită valoare dată.

Principiul curbei normale ne este folositor în multe feluri. El este necesar chiar pentru testarea ipotezelor, dar acest subiect este tratat într-un alt modul de curs.

Întrebări pentru studiu

1. Cum diferă o distribuţie oblic pozitivă faţă de o distribuţie oblic negativă?
2. Care sunt caracteristicile unui poligon de frecvenţe normal, sau sub formă de clopot?
3. Cum arată distribuţia într-un poligon de frecvenţe pentru variabila "număr de copii în familiile cu ajutor social", normală, oblic pozitivă, sau oblic negativă?
4. Într-o distribuţie oblic pozitivă, unde se află mediana faţă de medie?

5. La o variabilă care este normal distribuită, aproximativ ce procent din toate valorile se găsesc la o distanță de o abatere standard față de medie?
6. Ce valoare Z corespunde unei anxietăți de 79 puncte când media eşantionului de date este de 84, iar abaterea standard este 5?
7. Într-o distribuție normală, cât de frecvent se poate întâmpla să se găsească o valoare care este mai mare decât trei abateri standard peste sau sub medie?
8. La un test IQ cu media de 100 şi abaterea standard 16, la ce centilă va cădea un IQ de 132?
9. Ce valoare Z reflectă o valoare brută mai mare, -1,62 sau +1,50 ?

Exercitiu pentru acasă 6

Dacă o femeie are o valoare Z de -0,71 pentru greutate şi 1,95 pentru înălțime, pe o scală normativă determinată pentru toate femeile din țara, cum pare să fie corpul ei față de celelalte?

Modul 7

Introducere in SPSS

7.1. Procesul de analiză statistică a datelor în SPSS

În principal, SPSS a fost conceput pentru ajuta utilizatorul în ducerea la capăt a unei cercetări. Desigur, în procesul cercetării analiza statistică intervine mult mai târziu. Înainte de a ne aşeza în faţa unui calculator şi a folosi Editorul de date (modulul Data Editor) al SPSS pentru a introduce datele, se presupune că am parcurs deja paşi importanţi în procesul unei cercetări.

Computerul şi pachetul de programe SPSS intervine doar în partea de analiză a datelor. În faţa calculatorului activitatea se desfăşoară practic în patru paşi care, de obicei, se reiau de mai multe ori (vezi figura 7.1).

Pasul care ne interesează în prima fază este cel de introducere a datelor într-un fişier de date sau, după caz, de aducere a datelor în memoria internă.

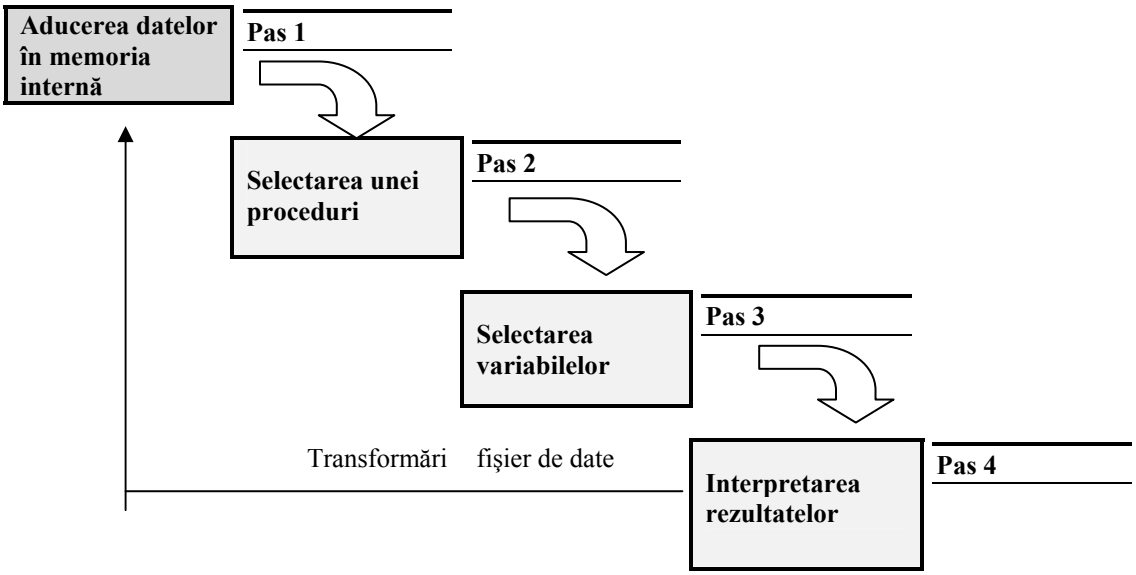


Figura 7.2. Procesul de analiză a datelor

1. Aducerea datelor în memoria internă

- Datele de gestionat (sau, după caz, de analizat) pot fi date salvate anterior într-un fişier de date SPSS.

- Se pot citi date din fişiere proprii altor produse software. Cele mai cunoscute sunt bazele de date dBASE, FoxPro sau Access; fişierele tip foaie de calcul produse cu tabelatorul Excel; sau fişiere simple de tip text ASCII, produse cu orice editoare de texte. Desigur, pentru a "importa" astfel de date, trebuie respectate nişte reguli clare şi trebuie furnizate informaţii suplimentare de conversie, în cadrul procedurilor iniţiate de SPSS.

- Se pot introduce direct datele în formatul propriu fişierelor de date SPSS cu ajutorul modulului Data Editor.

2. Selectarea unei proceduri

Sistemul de meniuri al SPSS oferă o gamă de comenzi care permit selectarea şi declanşarea a patru categorii de proceduri:

- Proceduri de gestionare a datelor
- Proceduri de obținere a listelor şi rapoartelor
- Proceduri de construire a tabelelor statistice, de calcul a unor indicatori statistici, de căutare a unor modele explicative
- Proceduri pentru crearea unor grafice

3. Selectarea variabilelor

Cele mai multe din procedurile SPSS cer precizarea variabilelor. Acestea sunt listate în casetele de dialog ale procedurilor, iar utilizatorul va trebui să indice variabilele care intră în analiză.

4. Interpretarea rezultatelor

După ce toate elementele de informații necesare execuției unei proceduri au fost precizate de către utilizator, procedura se va declanşa şi va produce rezultatele. Ele trebuie privite cu atenție şi interpretate.

O şedință de lucru cu SPSS constă în mai multe reluări a celor patru faze: în mai multe operațiuni de management al datelor, sau mai multe proceduri de analiză a datelor. De multe ori ceea ce hotărâm să facem în continuare depinde de rezultatele unei operațiuni anterioare. Utilizatorul (sau după caz, cercetătorul) foloseşte SPSS ca un asistent docil, dar foarte puternic şi capabil, ale cărui performanțe depind de abilitatea cercetătorului de a lua deciziile cele mai potrivite.

SPSS sub Windows funcționează sub controlul unui mediu grafic GUI (Graphical User Interface), utilizând un sistem de meniuri descriptive şi casete de dialog care uşurează mult interfața om-calculator. Cele mai multe operațiuni sunt declanşate şi duse la capăt prin punctare şi clic cu mouse-ul, chiar dacă lucrul cu tastatura este oricând o a doua alternativă.

Module si ferestre SPSS

Ca orice produs sub Windows, SPSS se porneşte prin dublu-clic pe pictograma sa aflată fie pe ecranul dispay-ului, fie în lista de *Programe* al meniului *Start*.

Ceea ce vom vedea prima dată va fi modulul Data Editor (vezi figura 1.2). În afara acestui modul, SPSS mai are încă două: modulul Output Navigator – modul care se ocupă de afişarea rezultatelor prelucrărilor, precum şi de modificarea aspectului lor - şi modulul Chart Editor – modul care facilitează manipularea şi gestionarea graficelor obținute prin SPSS.

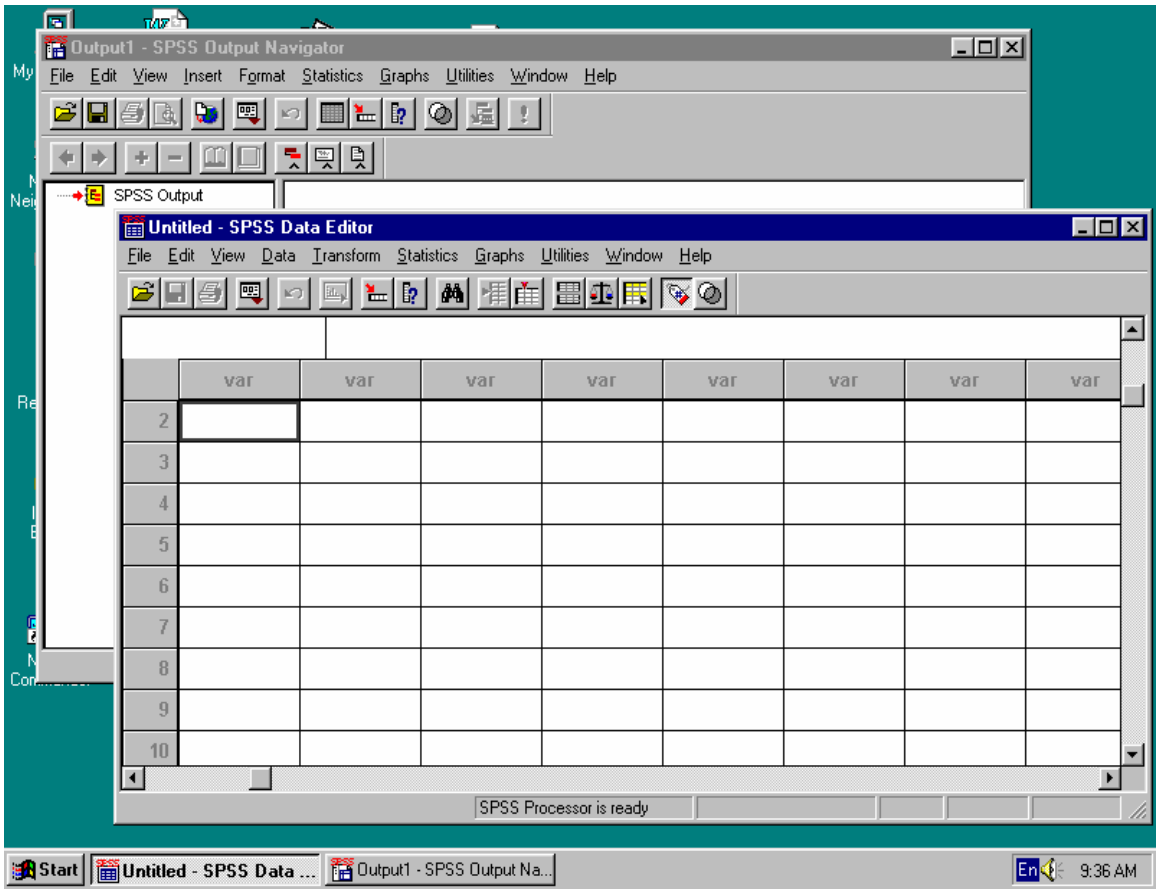

Output1 - SPSS Output Navigator
File Edit View Insert Format Statistics Graphs Utilities Window Help
SPSS Output
Untitled - SPSS Data Editor
File Edit View Data Transform Statistics Graphs Utilities Window Help
var
SPSS Processor is ready
Start
Untitled - SPSS Data ...
Output1 - SPSS Output Na...
9:36 AM

Figura 7.2. Ferestrele Data Editor şi Output Navigator

Fiecare fereastră are o linie de titlu. Sub ea se găseşte meniul principal al aplicaţiei. Fereastra Editorului de date conţine datele de analizat sub forma unui tabel asemănător unei foi de calcul, iar orice rezultat produs de o procedură SPSS se va introduce în fereastra Output al Navigatorului de rezultate. Iniţial se va deschide doar fereastra Editorului de date. Fereastra Navigatorului de rezultate se deschide îndată ce vreo procedură produce rezultate. In afara acestor două ferestre mai pot apare, la nevoie, alte două ferestre: fereastra Editorului de grafice sau fereastra Sintax (pentru păstrarea comenzilor SPSS în vederea repetării lor într-o altă şedinţă de lucru).

Sistemul de meniuri

SPSS v7.5 poate lucra într-o varietate de moduri, dar majoritatea procedurilor pot fi accesate făcând selecţii din meniuri. Meniul principal al Editorului de date, cea mai utilizată componentă a pachetului, conţine nouă meniuri:

File. Se utilizează pentru a crea fişiere noi SPSS, pentru a deschide fişiere existente, pentru a citi fişiere de tip spreadsheet sau bază de date create cu alte programe software. Meniul File se mai utilizează pentru crearea şi deschiderea altor fişiere SPSS, adică fişiere de grafice, de rezultate sau de sintaxă.

Edit. Se utilizează pentru a modifica sau copia texte din ferestrele de rezultate sau de sintaxă.

View. Se utilizează pentru a modifica aspectul ecranului în funcţie de opţiunile utilizatorului.

Data. Se utilizează pentru a face schimbări globale la fişierul de date, cum ar fi agregarea datelor din mai multe fişiere, împărţirea cazurilor în subseturi. Aceste schimbări sunt doar temporare şi nu afectează fişierul permanent cu excepţia cazului când se cere explicit acest lucru (prin salvare).

Transform. Se utilizează pentru a face schimbări variabilelor selectate din fişierul de date şi pentru a calcula variabile noi pe baza valorilor existente în alte variabile. Nici aceste schimbări nu afectează fişierul permanent decât dacă schimbările se salvează în mod explicit.

Statistics. Procedurile statistice se selectează din acest meniu. Cele mai comune proceduri statistice sunt: tabelele de frecvenţe, tabelele de asociere, calculul indicatorilor statistici descriptivi, analiza varianţei, corelaţia sau regresia liniară.

Graphs. Meniul Graphs se foloseşte pentru a crea diagrame de bare, diagrame circulare, histograme, diagrame de împrăştiere şi alte grafice în culori şi de mare rezoluţie. Chiar şi unele proceduri statistice pot genera grafice. Orice grafic poate fi îmbunătăţit cu ajutorul Editorului de grafice (Chart Editor).

Utilities. Se foloseşte pentru a afişa informaţii despre variabilele din structura fişierului de date, pentru a defini şi utiliza seturi restrânse de variabile, sau pentru a deschide un index al comenzilor SPSS.

Window. Se utilizează pentru a aranja, a selecta şi a controla atributele diverselor ferestre SPSS.

Help. Acesta deschide ferestre de asistenţă "on-line" cu manuale electronice, demonstraţii, sfaturi contextuale şi programe de învăţare asistată de calculator de tip "tutorial".

Linia de pictograme

Cele mai comune comenzi SPSS, folosite în mod uzual, pot fi accesate prin intermediul unor pictograme situate imediat sub linia de meniuri, în linia de pictograme (Tool Bar). Această linie este prezentă la oricare din ferestrele modulelor SPSS, dar poate conţine pictograme diferite, adaptate situaţiei (vezi figura 7.2).

Unele din pictograme s-ar putea întâmpla să nu fie eligibile în anumite situaţii chiar dacă apar pe ecran. O scurtă explicaţie a ceea ce reprezintă ele poate fi obţinută foarte simplu prin plasarea indicatorului mouse-ului pe ele.

Casetele de dialog

Cele mai multe proceduri SPSS afişează casete de dialog prin intermediul cărora adună toate informaţiile necesare ducerii la bun sfârşit a operaţiunilor cerute de utilizator.

Pentru că aceste casete oferă informaţii utile în alegerile pe care utilizatorul le are de făcut şi pentru că aceste casete sunt foarte asemănătoare între ele, să aruncăm o privire, de exemplu, pe cea folosită la deschiderea unui fişier: figura 7.3

Această casetă conţine o zonă etichetată Look in unde va apare directorul curent (BazeDate, în exemplu). Dacă fişierul căutat nu este în directorul curent, acesta poate fi schimbat cu ajutorul pictogramelor din această zonă.

Fişierele sunt listate în zona cea mai întinsă a casetei de dialog. De exemplu, în figura 2.2 putem vedea patru fişiere. Oricare dintre ele poate fi deschisă. Alegerea se face prin pointare şi clic pe butonul stâng al mouse-ului.

Numele fişierului selectat va apare in zona File name. Această informaţie poate fi şi tastată dacă se doreşte folosirea tastaturii şi nu a mouse-ului.

Nu toate fişierele din directorul curent sunt listate în casetă. Ele sunt filtrate cu ajutorul extensiilor consacrate pentru diverse tipuri de fişiere. Observăm că în exemplu, toate cele trei fişiere au extensia .sav. Zona Files of type oferă însă posibilitatea selectării altor tipuri inclusiv a tuturor tipurilor, deci posibilitatea neutilizării filtrelor.

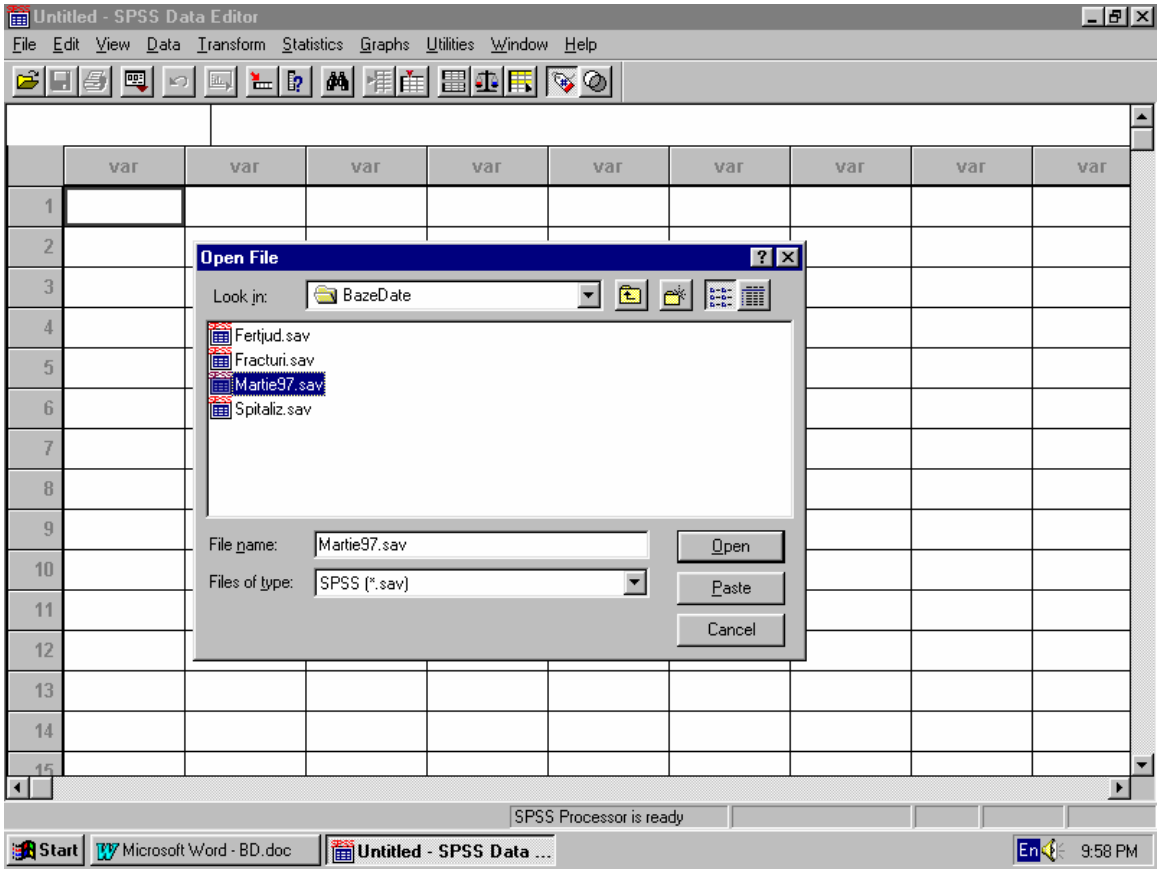


Figura 7.3. Caseta de dialog Open File

În partea dreapta jos a casetei se găsesc "butoanele", adică acele zone care declanşează acţiuni. Clic pe butonul Open produce deschiderea fişierului selectat. Clic pe butonul Paste produce doar un text cu comanda de deschidere a fişierului selectat, comandă ce se va scrie într-un fişier de tip sintaxă, în vederea folosirii lui ulterioare. Butonul Cancel anulează operaţiunea în curs, fără să declanşeze procedura pentru care se făceau pregătirile. După acţionarea butoanelor casetele de dialog dispar de pe ecran.

În majoritatea casetelor de dialog apare şi butonul Reset. Selectarea lui va anula toate specificările făcute în caseta de dialog afişată, fără ca acesta să fie ştearsă de pe ecran. Utilizatorul va putea face deci alte selecţii.

Unele casete de dialog pot conţine "butoane radio" şi / sau "căsuţe pentru bifare" (check box). Ele nu declanşează acţiuni imediate dar permit optarea pentru anumite variante ale procedurilor. Butoanele radio se deosebesc de căsuţele pentru bifare prin faptul că numai una dintre opţiuni poate fi aleasă şi una din alegeri este obligatorie. În cazul căsuţelor pentru bifare, acestea pot fi bifate sau nu, indiferent câte sunt într-o casetă de dialog. Butoanele radio au o aparenţă circulară iar opţiunea aleasă se marchează cu un punct (•). Căsuţele pentru bifare au aparenţă pătratică şi opţiunile alese vor fi marcate cu [4] .

Asistența "on-line"

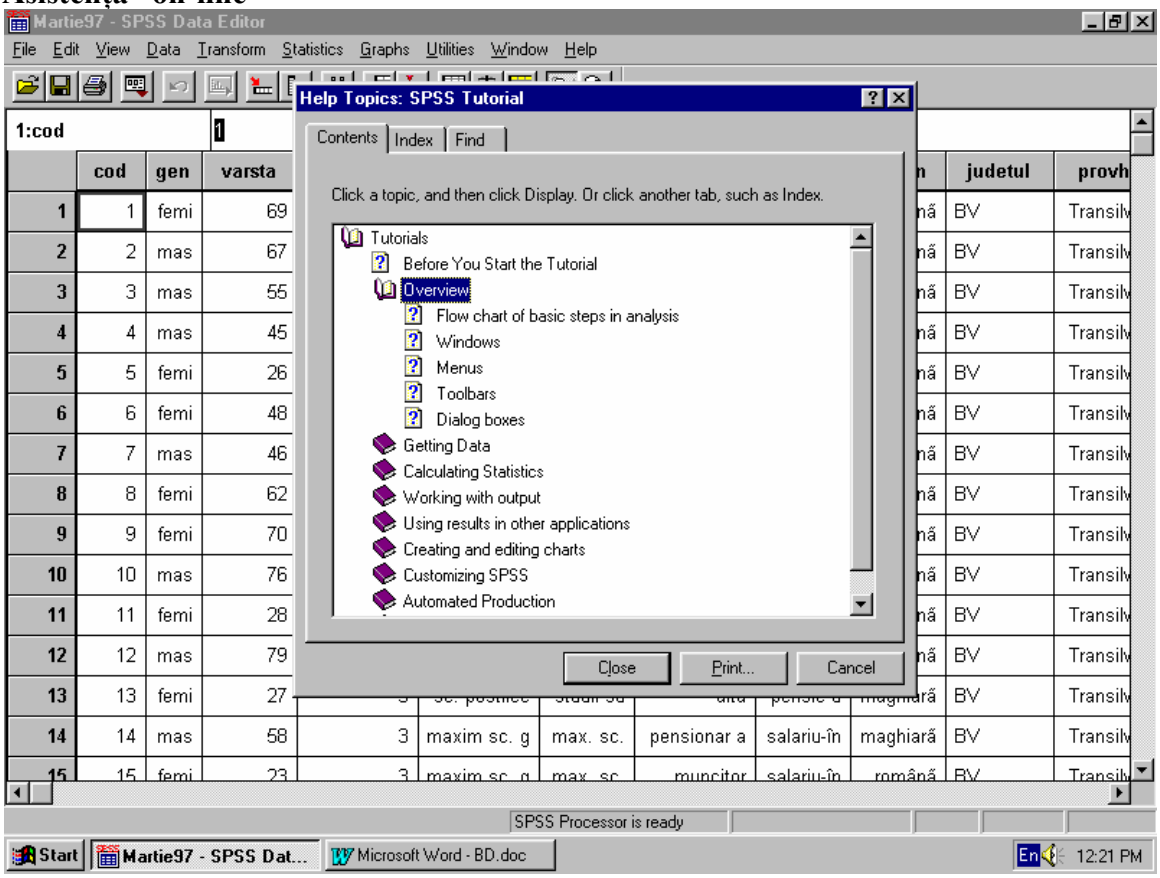


Figura 7.4. Fereastra Tutorialului SPSS

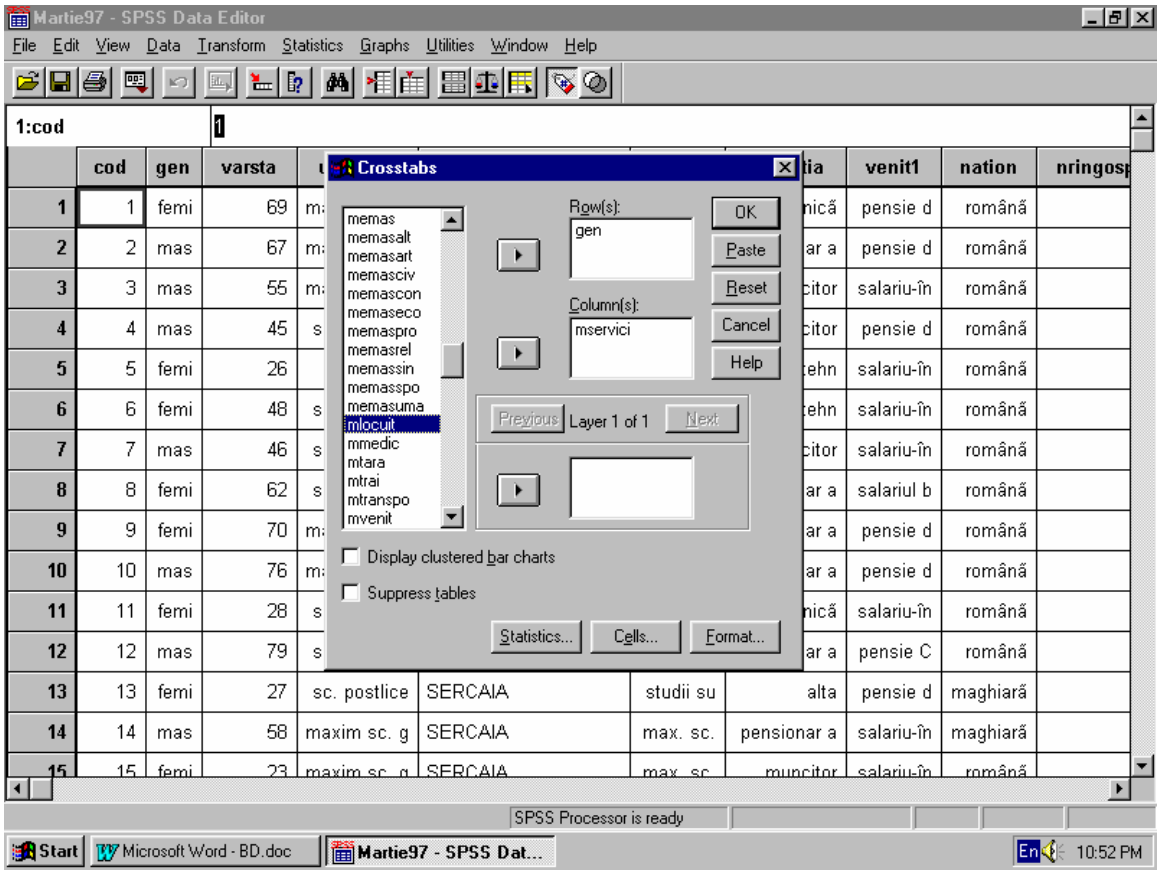
Meniul Help oferă multiple posibilități de a obține ajutor interactiv. Procedura Topics permite accesul în manualul electronic pe baza unor opțiuni tematice, procedura Tutorial este indicată pentru un prin contact cu SPSS (vezi figura 7.4), iar procedura Statistics coach este un meditator pentru procedurile statistice. Unele din proceduri conțin demonstrații interactive, cele mai multe conțin ilustrații și toate conțin explicații textuale în limba engleză.

Vom explora în cele ce urmează câteva din cele mai comune căi de a descrie datele dintr-o bază de date prin intermediul procedurilor de statistică elementară Frequencies și Descriptives. De foarte multe ori vom vedea că aceste două proceduri sunt suficiente pentru a răspunde la multe din întrebările cercetării. Procedurile Crosstabs și Means sunt utile în investigarea unor posibile relații între două variabile. Aceste patru proceduri, la care se adaugă și procedura Explore, nu sunt numai tehnici puternice descriptive, dar constituie tot atâtea mijloace de investigație necesare înaintea întreprinderii unor analize statistice mai sofisticate, de testare a ipotezelor.

7.2. Procedura Frequencies

Procedura *Frequencies*, furnizează tabele statistice și reprezentări grafice. Implicit, ea va genera o distribuție de frecvențe sub formă de tabel, în care pentru fiecare valoare distinctă a variabilei selectate, se va afișa numărul de apariții, procentul pe care acesta-l reprezintă din întregul eșantion și procentul cumulativ. Ca opțiuni, pot fi selectate grafice ca diagrama de bare, histograma (care vor fi alese în funcție de tipul variabilei), valorile pentru anumite percentile sau indicatori statistici descriptivi.
În tabela de frecvențe sau în diagramele de bare, valorile distincte pot fi ordonate în ordine crescătoare sau descrescătoare. Generarea / afișarea tabelei de frecvențe poate fi eliminată dacă variabila are prea multe

valori distincte. Unităţile de măsură folosite în diagrame şi grafice pot fi sub formă de frecvenţe numerice (implicit) sau procente.



Indicatori statistici ce pot fi ceruţi: media, mediana, modul, suma, abaterea standard, varianţa, amplitudinea, minimum şi maximum, eroarea standard a mediei, oblicitatea (skewness) şi ascuţimea (kurtosis) (cu erorile lor standard), cuartile şi percentile definite de utilizator, frecvenţe, procente, procente cumulate.
Pentru declanşarea procedurii se va alege din meniu: *Statistics, Summarize, Frequencies*

7.3. Procedura Crosstabs

În cazul în care dorim să studiem o relaţie de asociere între variabile categoriale, va fi utilă o tabelă care conţine frecvenţele combinaţiilor valorilor (categoriilor) celor două variabile. Acestea sunt tot tabele de frecvenţe, dar se generează nu doar pentru o variabilă, ci pentru combinaţia a două variabile. Vom numi o astfel de tabelă, *tabelă de frecvenţe între două variabile, tabelă de asociere* sau *tabelă de contingenţă*. Ea se obţine din meniu, alegând:

Statistics, Summarize, Crosstabs…

Pentru exemplificare, vom utiliza fişierul *Martie97.sav*, care conţine variabilele *gen* şi *mservici* (satisfacţia în muncă). Putem studia relaţia între *mservici* şi *venit*, sau între orice alte variabile între care bănuim că ar putea exista o legătură. Conţinutul celulelor aşa cum apare sub formă de frecvenţe absolute nu exprimă relaţia dintre cele două variabile. Pentru a putea compara diversele grupuri între ele, trebuie să folosim o măsură independentă de numărul cazurilor cuprinse în grupul respectiv; fiecare grup având un număr diferit de cazuri (359 persoane de sex masculin şi 293 persoane de sex feminin). O exprimare în procente a frecvenţelor din fiecare celulă permite comparaţiile dorite.

Genul respondentului * Multumit de - serviciu Crosstabulation

Count

		Multumit de - serviciu				
		deloc multumit	nu prea multumit	destul de multumit	foarte multumit	Total
Genul respondentului	masculin	32	61	190	76	359
	feminin	40	60	148	45	293
Total		72	121	338	121	652

Tabel de asociere între gen şi satisfacţia în muncă (frecvenţe absolute)

Genul respondentului * Multumit de - serviciu Crosstabulation

			Multumit de - serviciu				
			deloc multumit	nu prea multumit	destul de multumit	foarte multumit	Total
Genul respondentului	masculin	Count	32	61	190	76	359
		% within Genul respondentului	8.9%	17.0%	52.9%	21.2%	100.0%
	feminin	Count	40	60	148	45	293
		% within Genul respondentului	13.7%	20.5%	50.5%	15.4%	100.0%
Total		Count	72	121	338	121	652
		% within Genul respondentului	11.0%	18.6%	51.8%	18.6%	100.0%

Tabel de asociere între gen şi satisfacţia în muncă (frecvenţe procentuale)

Procentajul pe coloane ne arată distribuţia variabilei de pe linii pentru fiecare categorie a variabilei de pe coloane (ce procent din totalul cazurilor de pe o coloană este distribuit pe fiecare linie). Tabelul din figura 7.4 nu afişează astfel de procente.

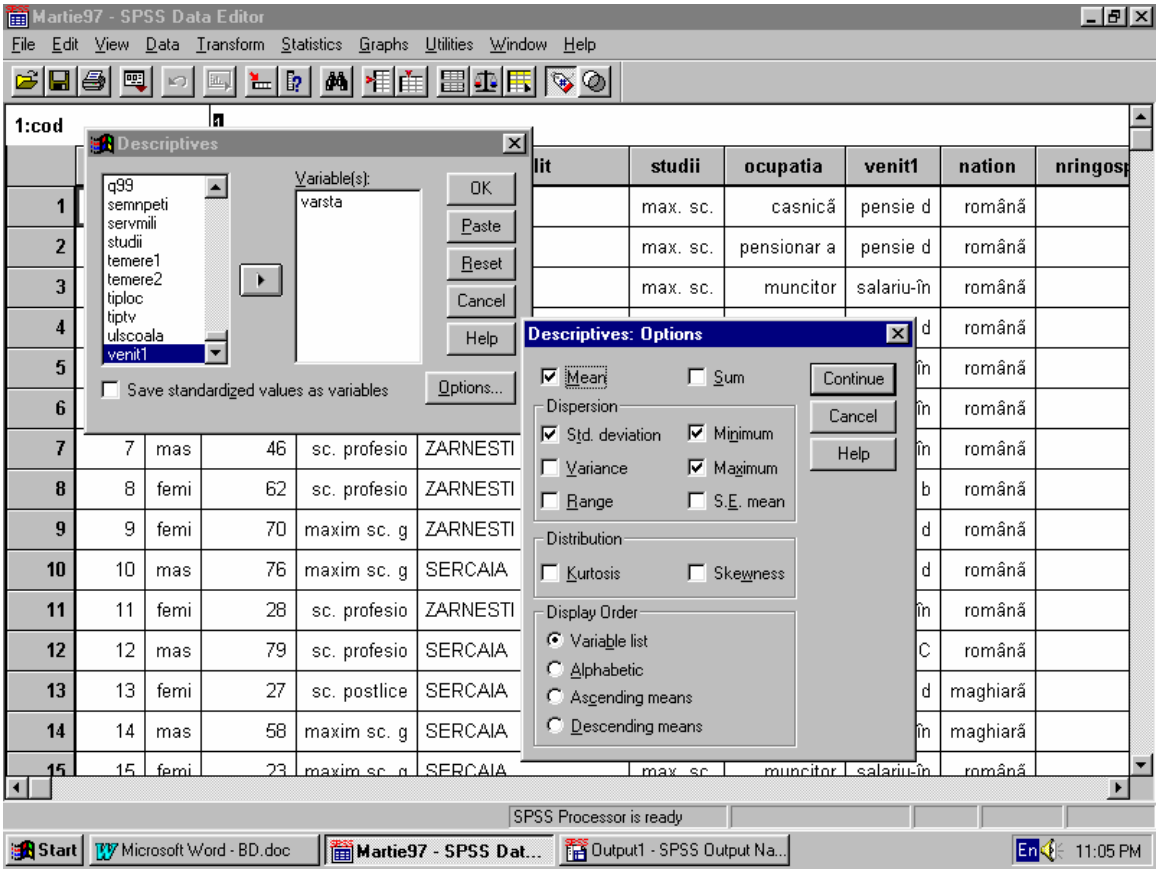
Procentajul pe linii ne arată distribuţia variabilei de pe coloane pentru fiecare categorie a variabilei de pe linii (ce procent din totalul de cazuri de pe o linie este distribuit în fiecare coloană). În tabelul din figura 7.4 putem vedea, de exemplu, că 8,9% din persoanele de sex masculin (variabila *gen* este variabila de pe linii) se declară deloc mulţumiţi de serviciul pe care îl au (variabila de pe coloane), 17% nu sunt prea mulţumiţi, 52,9% sunt mulţumiţi şi 21,2% sunt foarte mulţumiţi.

Observaţie: Cum putem spune dacă într-un astfel de tabel procentele din celule reprezintă procentajul pe linii sau pe coloane? Dacă în coloana numită *Total* procentul înscris în fiecare celulă este 100%, avem de-a face cu procente pe linii. Dacă pe linia numită *Total* procentul înscris în fiecare celulă este 100%, avem de-a face cu procente pe coloane.

7.4. Procedura Descriptives

Această procedură calculează şi afişează indicatori statistici univariaţi (o variabilă la un moment dat) pentru diferite variabile şi poate calcula valorile standardizate (scoruri *Z*) pentru fiecare caz. Implicit, pentru fiecare variabilă ce se selectează se vor calcula *media, abaterea standard, minimum, maximum* (vezi căsuţele bifate din caseta de dialog *Descriptives Options* din figura 7.5.). Opţional mai pot fi cerute varianţa, amplitudinea, eroarea standard a mediei, oblicitatea şi ascuţimea.

Observaţie: Mediana, modul, cuartilele şi percentilele nu pot fi calculate aici, ele se determină cu procedura *Frequencies.*



Procedura Descriptives

Declanşarea procedurii se face din linia de meniu cu:

Statistics, Summarize, Descriptives

7.5. Procedura Means

Într-un grup de subiecți putem distinge subgrupuri de cazuri în funcție de diverse criterii, subgrupuri între care este util să facem comparații în scopul stabilirii unor diferențe ce merită exploatate mai îndeaproape. Putem studia de exemplu media notelor la matematică obținute într-o clasă, separat pentru fete şi băieți. Sau putem studia diferențele privind media de vârstă pentru femei şi bărbați într-un eşantion reprezentativ al populației României (vezi figura).

Subgrupurile pot fi formate atunci când cazurile pot fi divizate pe baza uneia sau mai multor variabile de grupare.

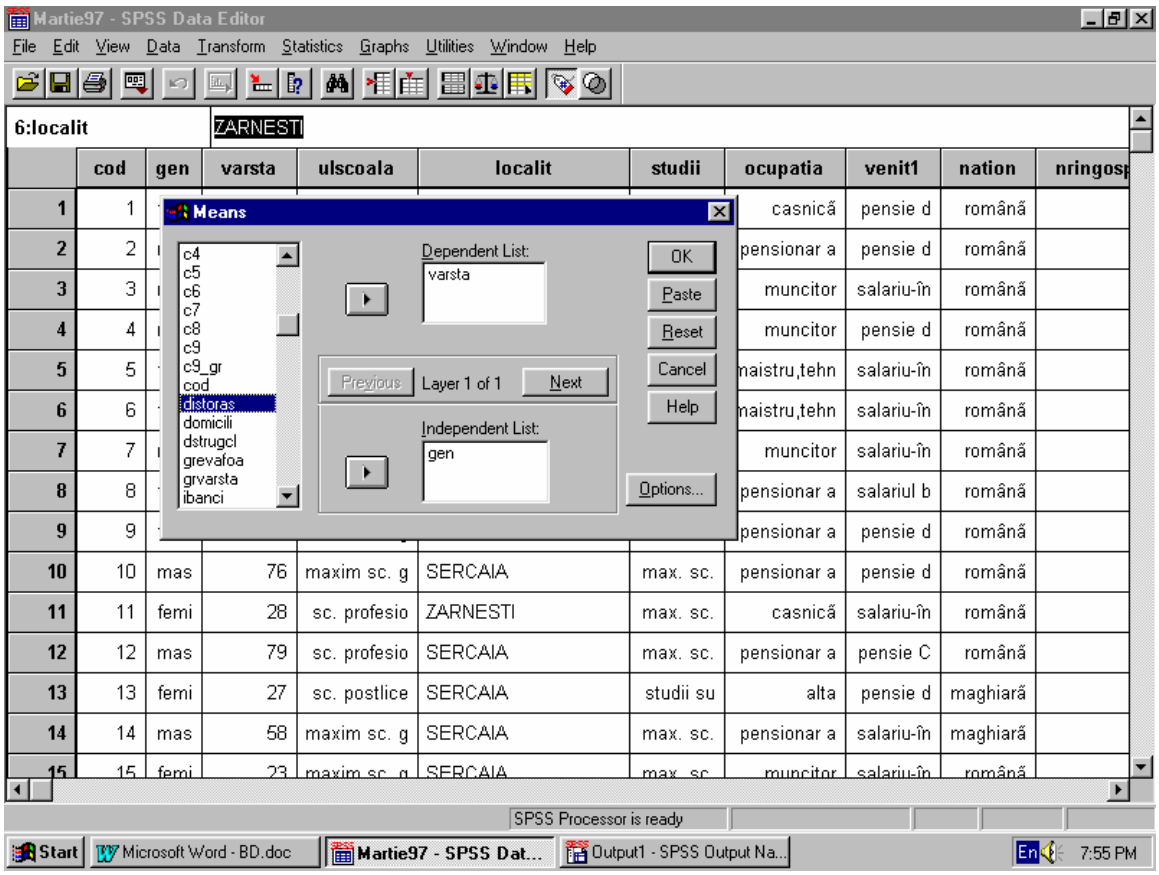
Variabila sub studiu este deci de tip interval iar variabila de grupare este categorială.

Declanşarea procedurii se face prin intermediul liniei de meniu:

Statistics, Compare Means*, Means*

Rezultatul este o tabelă care afişează pentru variabila în cauză media, abaterea standard şi efectivul fiecărei categorii de grupare.

În cazul în care se doreşte definirea grupurilor în funcție de mai multe variabile de grupare, variabilele categoriale vor fi puse în "straturi" (*Layers*) diferite.



Procedura Means

7.6. Procedura Explore

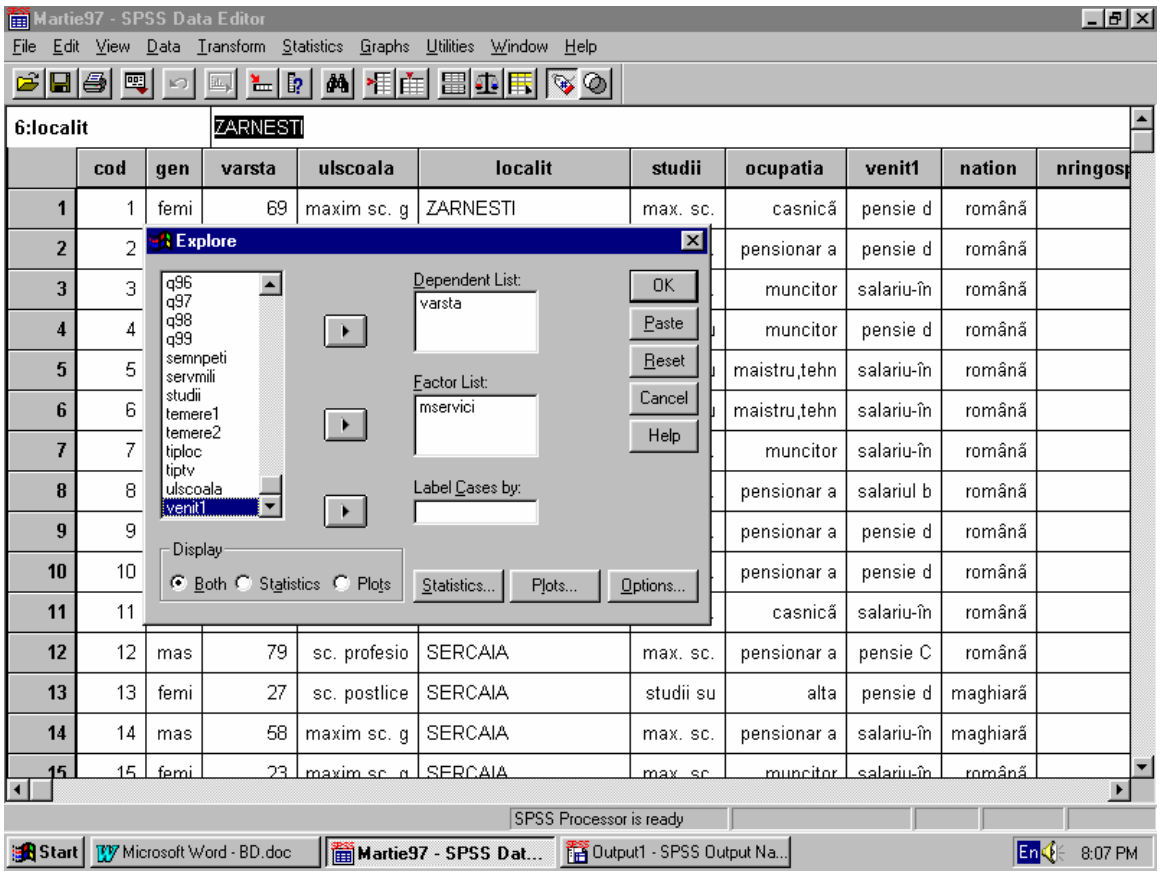
Poate fi folosită pentru a studia distribuția unei variabile:

- Calculează indicatori statistici descriptivi pentru toate cazurile sau pentru subgrupuri de cazuri.
- Identifică valorile extreme. Acestea se datorează uneori introducerii greşite a datelor în calculator. Dacă sunt reale, atunci ele pot influența puternic deciziile pentru analiza statistică ulterioară, de aceea este util să fie depistate.
- Calculează percentilele distribuției variabilei, atât pentru toate cazurile cât şi pentru subgrupuri de cazuri;
- Generează o varietate de reprezentări grafice a datelor (*stem and leaf, boxplot*, histograme) care ne arată vizual cum se distribuie valorile datelor.

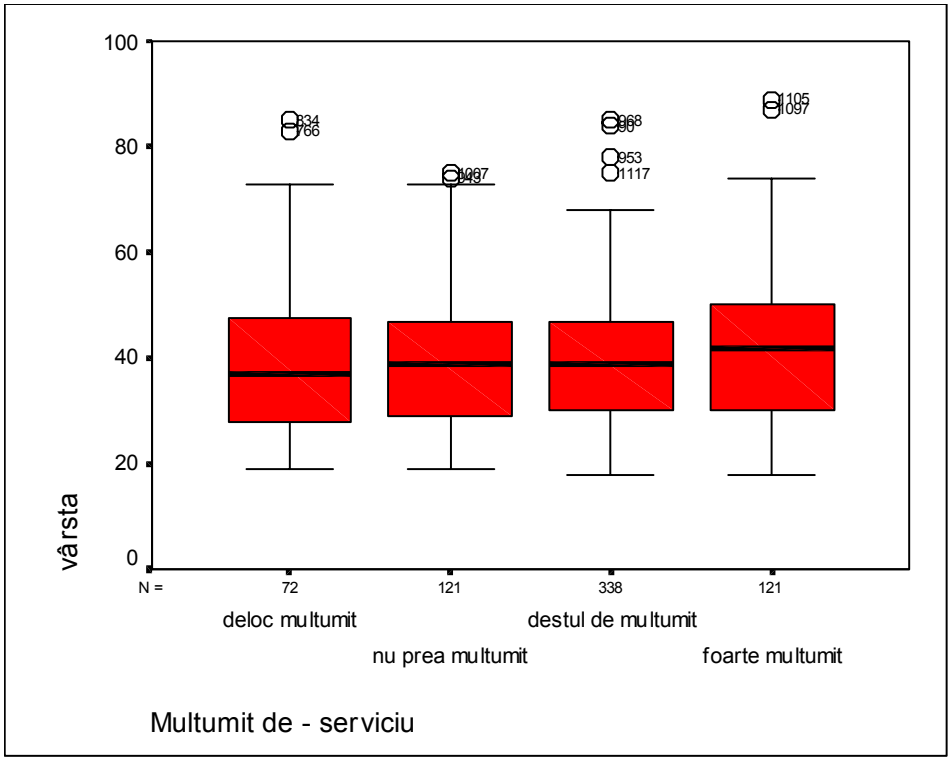
Pentru declanşare vom alege din linia de meniu:

Statistics, Summarize, Explore

În *Dependent List* se trece variabila pe care dorim s-o explorăm (variabilă tip interval). Variabila factor este o variabilă de grupare, care se alege în cazul în care dorim să calculăm statistici pentru subgrupuri de cazuri ale variabilei numerice studiate.



Caseta de dialog al procedurii Explore



Grafic boxplot pentru legătura dintre satisfacția în muncă şi vârstă (Martie97)

În figura de mai sus putem vedea o diagramă “boxplot” produsă de procedura *Explore* având ca şi variabilă factor *mservici* (satisfacția în muncă) iar ca variabilă dependentă *vârsta*. Caseta de dialog al procedurii *Explore* a fost completată ca în figura.

Modul 8.

Notiuni introductive de statistica inferentiala

Statistica descriptiva se ocupa de organizarea, tabelarea şi rezumarea datelor colectate despre caracteristicile unei populaţii statistice. Procedurile descriptive sunt destinate să prezinte datele într-o formă care să fie cât mai sugestivă pentru cei interesaţi de cunoaşterea empirica a realităţii. Deşi, pentru un observator atent, prezentările grafice şi indicatorii statistici descriptivi pot sugera legături între variabile, totuşi ele nu pot confirma şi nici verifica astfel de legături. Ele nu ne permit să generalizăm nici un fel de constatări dincolo de eşantioanele pe care le-am studiat. Pentru astfel de generalizări, avem nevoie să dezvoltăm ipoteze şi să le testăm în concordanţă cu principii ştiinţifice.

8.1. Testarea ipotezelor

Formularea ipotezelor în cercetarea şi practica de asistenţă socială este extrem de importantă. Fie că sunt validate sau nu, ipotezele ne permit să abordăm în mod ştiinţific practica de asistenţă socială.

Ipotezele sunt formulate deobicei după trecerea în revistă a literaturii de specialitate sau prin procesul sintetizării cunoştiinţelor cantitative şi calitative pentru o temă particulară: din experienţa profesională câştigată în practică; din cursuri scrise de profesionişti şi din multe alte surse specifice domeniului, incluzând legislaţia, documente nepublicate şi persoanele "cunoscătoare". După o examinare a literaturii, putem expune doar răspunsuri parţiale sau putem finisa răspunsuri la întrebări generale. La început se doreşte exprimarea impresiilor sau a concluziilor noastre sub forma ipotezelor. După aceea ipotezele vor trebui testate.

Multe definiţii au fost date ipotezelor, dar toate au la bază acelaşi concept - o ipoteză este o tentativă de răspuns la o întrebăre de cercetat, derivată din trecerea în revistă a literaturii de specialitate sau inspirată din practica curentă. Este în acelaşi timp o exprimare a unei relaţii dintre două sau mai multe variabile. O ipoteză, fie că este sau nu susţinută de analiza datelor, ea trebuie să fie o extensie logică a cunoştiinţelor existente anterior. Oricare altă metodă folosită pentru formularea ipotezelor nu este altceva decât "smulgerea" unei legături şi generează posibilitatea de pune la îndoială credibilitatea studiului de cercetare.

Dupa colectarea, organizarea şi rezumarea informaţiilor, utilizând statistici asemănătoare celor prezentate în prima parte a cursului, putem începe să înţelegem dacă ipotezele noastre sunt sau nu susţinute de datele colectate. Să presupunem, de exemplu, că am lansat ipoteza: clienţii alcoolici care primesc tratamentul în grup se pot abţine de la băutură mult mai bine, după trei luni de tratament, decât clienţii alcoolici care au primit tratamentul individual. Frecvenţele şi procentajele par să ne furnizeze suportul pentru aceată ipoteză; mai ales că numărul clienţilor a fost mic şi ne-a fost uşor să obsevăm tendinţa de succes pentru o metodă de tratament (aici tratamentul în grup) în opoziţie cu altele (tratamentul individual). Probabil, 70% din clienţii alcoolici care au primit tratamentul în grup se pot abţine acum de la băutură în comparaţie cu 65% din cei care au primit tratamentul individual. Există "în aparenţă" o legătură între variabila dependentă (băutor / nebăutor) şi variabila independentă (tratamentul de grup / tratamentul individual). Cu toate acestea, doar un naiv ar concluziona, doar pe baza celor prezentate anterior, că tratamentul în grup este cea mai bună metodă de lucru cu alcolicii. Pentru a testa ipotezele noastre în mod ştiinţific, trebuie să examinăm şi alte explicaţii posibile.

8.2. Explicatii alternative

Aşa cum am afirmat, ipotezele noastre trebuie să fie sprijinite de date colectate. Dar e încă prematur să concluzionăm că există o relaţie adevărată între cele două variabile. O relaţie poate fi verificată doar după epuizarea tuturor explicaţiilor alternative care ar putea justifica legătura. În cazul nostru cele trei explicaţii alternative majore sunt: (1) deformarea, (2) alte variabile şi (3) şansa. Primele două sunt în

principal legate de metodologia cercetării, iar cea de-a treia este direct legată de testarea statistică. O bună cunoaştere a metodelor de cercetare folosite în asistența socială ne va ajuta să minimalizăm efectele primelor două explicații alternative - iar această carte ne va ajuta cu eliminarea efectelor celei de a treia explicație alternativă.

Deformarea

Prima explicație alternativă este deformarea realității. Influențele conjuncturale sunt o sursă sistematică de distorsiune, care afectează calitatea datelor colectate. Ele pot conduce la rezultatele eronate şi apoi la tragerea unor concluzii eronate. Deformări conjucturale pot apare când datele se colectează într-un moment neprielnic sau când studiul este influențat de evenimente exterioare, de vreo tendință conştientă sau inconştientă a persoanelor care colectează datele şi care nu au o percepție corectă asupra lor. Dacă dintr-un motiv oarecare, sau combinație de motive, datele au fost deformate, variabilele noastre dependente şi independente pot apărea ca fiind legate, când de fapt ele nu sunt. Posibilitatea ca deformarea să poată explica o legătură aparentă între două variabile este minimizată deobicei prin folosirea cu mare grijă a tehnicilor de măsurare.

Alte variabile

A doua explicație alternativă este influența altor variabile. Factorii care influențează variabila independentă pot deasemenea explica diferențele din cadrul variabilei dependente (a bea / a nu bea, în exemplul nostru). În exemplul nostru, aceştia ar putea fi intensitățile diferite ale suportului familiar pentru metoda de tratament şi nivele de calificare diferite ale asistenților sociali care practică cele douămetode. Un plan experimental bun, construit pe baze aleatoare, poate garanta un anumit nivel de control pentru celelalte variabile, dar planurile experimentale bune sunt rare în cercetarea în asistență socială. Alegerea unei tehnici metodologice corespunzătoare ne poate garanta că alte variabile nu sunt explicații reale ale unei aparente legături între două variabile. Căile prin care planurile de cercetare pot ajuta în controlul altor variabile sunt prezentate în general în textele privind metodologia cercetării.

Şansa

Cea de-a treia explicație alternativă este şansa, care poate fi referită şi prin termeni ca probabilitate, eroare de eşantionare, întâmplare norocoasă - sau pur şi simplu noroc. Şansa stipulează ca probabilitatea să se întâmple un eveniment se poate situa oriunde între valorile 0 (niciodată) şi 1 (absolut sigur). Ea se bazează pe presupunerea că în timp ce prin repetarea în timp îndelungat a unor observații se poate vedea ca există o anumită regularitate (pattern), în particular, sau pe termen scurt, observațiile tind să difere într-un fel de pattern-ul pe termen lung. De exemplu, când desemnăm în mod aleator, un eşantion dintr-o populație, putem să nu selectăm un eşantion care să aibe o compoziție identică cu cea a populației din care provine. Teoria probabilității ne trimite direct la conceptul de eroare de eşantionare, prezentat în textele metodologice.

Probabilitatea spune că o monedă lansată în aer, are 0,5 şanse (sau 50%) să cadă cu o anumita față în sus. Cu alte cuvinte, probabilitățile ca să fie cap sau pajură sunt egale. În realitate dacă aruncăm o monedă de zece ori am putea găsi un rezultat diferit de 5 cu 5. Nu vom fi surprinşi dacă obținem 4 cu 6, sau 8 fețe din zece aruncări. În acest caz vom învinovăți şansa (eroarea normală de eşantionare), şi vom presupune că dacă vom repeta testul de suficient de multe ori, procentul de aparinție al feței va fi aproximativ de 50%.

Când analizăm datele noastre, trebuie să determinăm dacă nu cumva rezultatele noastre sunt aberații plauzibile de la patternul normal al evenimentelor apărute din cauza erorilor de eşantionare. Trebuie să fim cât mai siguri că o legătură aparentă dintre variabile nu este un simplu noroc ce poate apărea din cînd în cînd. Avem nevoaie să determinăm dacă şansa poate fi explicația reală pentru o legătură aparentă. O ipoteză poate fi susținută doar când şansa, la fel ca şi deformarea şi ca influența altor variabile, pot fi convingător eliminate ca explicații alternative. Şansa este de obicei ultima explicație pe care consumatorii sceptici de rapoarte de cercetare îl propun ca şi cauză *reală* a legăturii aparente dintre două variabile. Spre deosebire de cazul deformării sau al altor variabile, planurile solide de cercetare nu sunt suficiente pentru a elimina şansa ca explicatie alternativă pentru o aparentă legătură între două variabile. Doar testele statistice o pot face.

8.3. Respingerea explicației șansei prin testele statistice

Toate testele statistice încearcă să discrediteze șansa ca explicație a unei aparente legături între variabile. Cu toate acestea, ele o fac pe căi diferite. Din fericire, între testele statistice sunt mai multe asemănări decât diferențe. Acest capitol examinează căile prin care toate testele statistice se apropie de acest obiectiv.

În încercarea de a câștiga suport pentru ipoteze, nu putem elimina total șansa ca explicație pentru o legătură aparentă. Înainte de susține o legătură, trebuie să ne asigurăm într-o măsură *rezonabilă* că cea ce am observat nu este o întîmplare norocoasă care poate fi ușor explicată prin eroarea normală de eșantionare derivată din întâmplare. Nu dorim să raportăm o legătură care pare să fie adevărată, dacă ea nu este de fapt așa. În același timp, nu trebuie să fim niciodată atât de rigizi sau neraționali, încât să nu cerem suport statistic pentru un rezultat al cercetării care este *foarte improbabil* să se datoreze șansei. Dacă cercetătorii devin obsedați de eliminarea totală a șansei ca explicație alternativă, puține, sau chiar niciunul din rezultatele găsite ar vedea lumina zilei.

Testele statistice determină probabilitatea ca relațiile aparente dintre variabile să se datoreze întâmplării. Dacă probabilitatea efectului întâmplării este mică, și dacă deformările și celelalte variabile au fost eliminate ca și explicații posibile pentru o legătură aparentă, rămâne o singură explicație rațională: există o legătură *credibilă*. Dacă efectul șansei este mare, noi nu mai putem pretinde suport pentru o ipoteză care prezice o legătură, chiar dacă celelalte două explicații alternative au fost eliminate. Pe scurt trebuie eliminate toate cele trei explicații alternative înainte ca o ipoteză să poată fi considerată ca având suport.

Tipuri de legături declarate în ipoteze

O ipoteză care declară o legătură între variabile, dar care nu indică care valori ale unei variabile se vor grupa cu care valori ale altei variabile, este denumită ***ipoteză nedirecțională*** (sau *two-tailed*) O ipoteză care declară o legătură între două variabile *și* specifică modul (direcția) în care se crede că sunt legate, este denumită ***ipoteză direcțională*** (sau *one-tailed*).

Exemplul folosit mai devreme în acest capitol este o ipoteză direcțională. Ea precizează clar că rata înaltă a abstinenței la alcool va fi găsită printre clienții alcolici care au primit tratamentul în grup față de clienții alcolici care au primit tratamentul individual. O ipoteză nedirecțională pentru exemplul nostru, ar stabili că metoda de tratament folosită este legată de abstinență; dar nu ar putea prezice care metodă de tratament poate fi asociată cu rata ridicată sau scăzută de abstinență.

Există, firește, o a treia descriere posibilă a legăturii între două variabile - predicția că ele nu vor fi găsite legate (nu există nici o relație între ele). Această este numită ***ipoteză nulă***. Deși ipotezele cercetării tind în mod frecvent să fie direcționale, câteodată sunt folosite și ipotezele nedirecționale, dacă cunoștințele asupra subiectului sunt limitate. Căutarea suportului pentru a prezice că două variabile nu sunt legate (caz in care ipoteza cercetării este chiar ipoteza nulă) este rară, dar nu fără utilizare în practica și cercetarea din asistența socială. Cercetătorii care au căutat să contrazică ipoteza că o rasă ar fi superioară alteia din punct de vedere intelectual, au căutat suport statistic pentru ipoteza nulă (rasa și inteligența nu sunt legate). Similar practica curentă poate sugera că tratamentul individual este mai eficient decât tratamentul în grup pentru tratarea disfuncționalității sexuale a clienților. Din observațile noastre și din literatura de specialitate am putea totuși concluziona că nu se pot face diferențieri între metodele de tratament folosite în asistența socială. În acest caz am putea alege ipoteza nulă ca ipoteză pentru o astfel de cercetare: succesul în tratarea disfuncției sexuale *nu este legat* de metoda de tratament.

Ipoteza nulă și ipoteza cercetării

Cele trei forme ale ipotezei (direcțională, nedirecțională și nulă) sunt toate importante pentru noi în construcția ipotezei unei cercetări. Toate trei, dar în special ipoteza nulă, sunt extrem de importante în înțelegerea modului în care indicatorii statistici sunt folosiți în testarea ipotezelor. Formal, testele statistice iau in considerare ipoteza nulă, chiar dacă aceasta nu este ***ipoteza cercetării*** .

Trebuie amintit că ipoteza nulă este consistentă în raport cu șansa. Ea susține că două variabile sunt nelegate chiar dacă ele par a fi legate într-un set de date. Ea continuă să susțină că doar fluctuația normală a șansei, sub forma erorilor de eșantionare, este cea mai potrivită explicație pentru aparenta legătură dintre două variabile. Chiar și atunci când ipoteza nulă nu este folosită ca ipoteză a cercetării, și

ipotezele sunt fie direcționale fie nedirecționale, conceptul ipotezei nule joacă un rol important în testare. Pentru a confirma că două variabile sunt legate, trebuie să verificăm mai întîi că ele nu sunt nelegate. Altfel spus trebuie să demonstrăm că şansa (ca expresie a ipotezei nule) nu este explicație fericită a legăturii aparente.

Când o ipoteză este direcțională sau nondirecțională, există un fel de fantomă a ipotezei nule. Ipoteza nulă statuează că o legătură aparentă este de fapt un rezultat al şansei. Înseamnă că am extras un eşantion netipic de date şi că legătura aparentă pe care am observat-o în eşantion nu caracterizează în realitate populația din care am extras eşantionul. Pentru a respinge ipoteza nulă, avem nevoie să demonstrăm că şansa este o explicație nefericită pentru legătura aparentă observată şi că o legătura adevărată este o concluzie mult mai plauzibilă. Testele statistice ne permit să determinăn când putem întări o ipoteză direcțională sau non direcțională cu ajutorul unui suport statistic.

Erori de tipul I şi de tipul II

Două tipuri de erori pot fi făcute în interpretarea rezultatelor unei cercetări: tipul I şi tipul II. O eroare de tipul I este atunci când se respinge ipoteza nulă şi se concluzionează că există o legătură între două variabile, când de fapt nu există nici o legătură. O eroare de tipul II este atunci când încercarea de a respinge ipoteza nulă şi de a identifica o legătură adevărată între două variabile eşuează, de fapt există una. Cele două tipuri sunt comparate în Tabelul 5.1.

Tabel 8.1. Erori de tip I şi II

În realitate	**Decizia noastră**	
	Respingem ipoteza nulă	**Acceptăm ipoteza nulă**
Ipoteza nulă este falsă	Corect	Eroare de tip II ex: ”criminalul este nevinovat”
Ipoteza nulă este adevărată	Eroare de tip I ex: ”nevinovatul este criminal”	Corect

Unii din factorii care influențează probabilitatea producerii erorilor de tipul I sau II privesc deciziile privind planul de cercetare. Aceştia includ selectarea unui eşantion deformat, utilizând instrumente de colectare a datelor care sunt nevalabile şi/sau nesigure, şi care vor sfârşi prin imposibilitatea de a controla efectul altor variabile.

Tipul I şi II de erori pot de asemenea rezulta din folosirea testelor statistice nepotrivite. Dacă folosim un test statistic ce necesită condiții puternice care nu sunt îndeplinite, sau dacă utilizăm un test care necesită doar câteva condiții şi în realitate sunt îndeplinite condiții pentru un test mai puternic, pot apărea erori de tipul I sau II. În primul caz datele au fost tratate ca şi când ar poseda calități care de fapt le lipsesc, în ultimul caz, oportunitatea pentru o analiză mai exactă nu a fost folosită. Dacă nu e folosit testul statistic potrivit, poate apărea o legătură statistică numai datorită însuşirilor eronate pe care le-am atribuit datelor şi modului în care au fost colectate. Sau s-ar putea ca o legătură adevărată să rămână ascunsă.

Niciodată nu putem elimina în totalitate posibilitatea de comitere a erorilor în luarea deciziilor, deci nici în decizia de a respinge sau nu ipoteza nulă. De fapt dacă suntem prea prudenți să nu comitem erori de tipul I (să respingem în mod greşit ipoteza nulă), creşte posibilitatea comiterii tipului II de eroare (greşeala de a nu respinge ipoteza nulă). Analog, grija exagerată de a nu comite erori de tipul II duce la creşterea probabilității de a comite erori de tipul I. Cercetătorii trebuie să decidă până la urmă care eroare, de tipul I sau de tipul II, este mai acceptabilă pentru ei. Aceasta este o decizie de natură etică ce presupune cunoştințe din practica asistenței sociale şi despre consecințele uneia sau alteia din erori. Din fericire aşa cum vom prezenta mai departe, există convenții statistice care să ne ghideze în luarea deciziilor.

Într-un studiu de cercetare importanța acestor erori este, desigur, potențial gravă. De exemplu, profesioniştii în asistență socială nerecunoscând că folosirea greşită a testului statistic sau a unei erori metodologice a dus la un rezultat eronat din tipul I, pot concluziona în mod eronat că există o legătură între o metodă particulară de tratament şi rata mai mare de succes în tratament. Ei pot ajusta accesul la tratament pe baza acestui “fapt”. Sau ei pot reacționa la alte rezultate ale cercetării în care (din anumite motive) a fost comisă eroare de tipul II, şi care discreditează o metodă de tratament în realitate foarte buna, dar care a apărut că nu face parte dintre tratamentele eficiente. Tipul I şi tipul II de erori pot fi la fel de distructive

când aplicăm rezultatele cercetării la situaţiile practice de asistenţă socială. Amândouă ne pot conduce la concluzii greşite, pot face rău clienţilor asistaţi, sau duc la risipirea resurselor limitate ale agenţiei.

Chiar dacă studiile de cercetare sunt bine planificate şi dacă noi înţelegem şi aplicăm criteriile de selecţie a testului statistic potrivit, rămâne întotdeauna posibilitatea comiterii unei erori în tragerea concluziilor intr-o cercetare. Întotdeauna rămâne posibilitatea de-a ni se întâmpla ca unul din miliardele de eşantioane posibile să ne conducă la concluzii eronate privind legăturile dintre variabile în populaţia din care provine eşantionul. Nu putem exclude in totalitate nici existenţa unor mici erori metodologice care sa fi introdus deformarea sau existenţa altor variabile cu impact asupra variabilei cercetate dar ignorate în planul nostru de cercetare. Această vagă posibilitate nu trebuie, totuşi, să ne timoreze în a ne asuma riscuri rezonabile în interpretarea rezultatelor cercetării şi în aplicarea lor practică. În felul acesta putem face progrese şi putem deveni practicieni cu o bază de cunoştinţe ştiinţifice.

8.4. Dovezi suficiente şi semnificaţie

Dacă ne întoarcem la exemplul nostru cu cele două tipuri de tratamente şi la posibila legătură cu abstinenţa clienţilor alcoolici, putem observa că diferenţa de procente dintre rata abstinenţei pentru grupul care a primit tratamentul în grup (70%) şi rata abstinenţei pentru aceia care au primit tratamentul individual (65%), este de fapt destul de mică. Chiar dacă am fi avut 50 de clienţi pentru fiecare tratament, mulţi dintre noi nu pot spune dacă aceste 5 procente (70% - 65% = 5%) diferenţă pot fi suficiente pentru a respinge ipoteza nulă (ipoteza nulă în acest exemplu poate fi aceea că metoda de tratament şi efectele nu au legătură una cu alta). O diferenţă de 40 procente între două metode de tratament, ne-ar da mai multe speranţe în a obţine suport statistic pentru o asemenea ipoteză.

Din nefericire, în realitate datele arareori vorbesc atât de tranşant pentru a putea susţine sau respinge ipoteza nulă. Ele de obicei sunt astfel încât avem nevoie de testele statistice care să ne ajute în a decide dacă avem sau nu suport statistic pentru eliminarea ipotezei nule. Până la ce punct ne putem simţi liniştiţi când vrem să respingem ipoteza nulă şi să susţinem că avem suport statistic pentru ipoteza direcţională? Dacă ar fi posibil să repetăm un studiu de cercetare (numit *replicare* în terminologia cercetării) de 100 sau chiar de 200 de ori şi în fiecare studiu clienţii trataţi în grup ar avea o rată de abstinenţă alcoolică mai înaltă, am fi mai convinşi că ipoteza nulă poate fi respinsă. Din nefericire, în cercetarea de asistenţă socială nu prea avem posibilitatea de a repeta studiile de mai multe ori. Astfel că avem nevoie să obţinem nişte dovezi palpabile ale improbabilităţii ipotezei nule, cu ajutorul unui singur studiu. De la ce punct putem fi suficient de siguri că o legătură aparentă nu poate fi respinsă în mod rezonabil, datorită aportului şansei? Aici trebuie să apelăm la simţul comun şi la convenţii.

De-a lungul anilor, cele mai multe cercetări au situat la nivelul de 95% certitudine, punctul care este suficient de înalt pentru a putea avea încredere în eliminarea ipotezei nule. Altfel spus cercetătorii se vor simţi siguri în a concluziona că două variabile sunt legate (sau în relaţie) dacă analiza statistica sugerează că există mai puţin de 5% procente şanse de a face o greşeală, prin respingerea ipotezei nule. Acest nivel de risc în comiterea unei erori de tipul I este acceptabil în majoritatea studiilor de cercetare. Convenţia statistică declară că există suport pentru o ipoteză direcţională sau nondirecţională, dacă *probabilitatea de a face o eroare de tipul I* este mai mică decât 5% - se mai utilizează şi terminologia echivalentă: *nivel de semnificaţie* 0,05 , sau *coeficient de risc* 0,05, sau *nivel de încredere* 0,95. O decizie care respinge ipoteza nulă nu poate elimina definitiv şansa ca explicaţie posibilă a unei legături aparente, dar ea este un argument pentru a putea declara o legătura ca fiind *statistic semnificativă*.

Nu există nimic sacru în ceea ce priveşte nivelul de semnificaţie 0,05, dar el este cel mai des folosit pentru respingerea ipotezei nule. Decizia folosirii şi a altor nivele decât cel de 0,05 se justifică în funcţie de consecinţele care ar rezulta prin luarea unei decizii greşite. O cerere de dovezi mai tari, pentru acceptarea unei legături între variabile, cum ar fi cele de nivel 0,025 sau 0,01, poate fi dorită când este necesar să existe chiar mai puţine posibilităţi de a greşi respingând ipoteza nulă şi de a concluziona că există o legătură între două variabile, atunci când întâmplarea este o explicaţie reală pentru legătura observată dintre ele. Dacă utilizarea rezultatelor cercetării poate fi o problemă de viaţă şi de moarte, aşa cum e în cazul unui nou medicament, este nevoie de un nivel de probabilitate mai înalt şi mai multă precizie în eliminarea ipotezei nule, de exemplu putem folosi nivelul 0,001. Când acceptăm o legătură la un nivelul de semnificaţie mai mic decât 0,001, spunem că ea este *foarte semnificativă statistic*. Într-adevar, în acest caz, probabilitatea de a respinge ipoteza nulă în mod eronat este de numai 1 la 1000, adică probabiltatea ca legătura să fie adevarată este de 99,9%.

În studiile de cercetare în care importanța unei erori în respingerea ipotezei nule nu este fatală sau traumatică, putem considera că nivelul de semnificație 0,10 (10% probabilitate de a greşi) este acceptabil. La nivel de semnificație 0,10 există de două ori mai multe posibilități de a comite o eroare de tipul I (sub forma erorii de eşantionare) decât există la nivelul 0,05. Când se acceptă un nivel de semnificație mai mic decât pragul critic 0,10 spunem ca avem legături *slab semnificative statistic*. Pentru confirmarea unei astfel de legături se cere ca planul de cercetare să includă replicarea (repetarea cercetării). Dacă un nivel de 0,10 poate fi considerat ca insuficient pentru a dovedi o legătură, un caz tipic sau o serie de cazuri tipice sunt suficiente pentru a ajunge la concluzia că ipoteza nulă poate fi respinsă.

Deşi o anumită flexibilitate este permisă în selectarea pragurilor la care şansa este în mod acceptabil eliminată ca explicație pentru o legătură aparentă, alegerea unui nivel de încredere nu trebuie văzută ca şi cauzală. Selectarea unui nivel de încredere trebuie deasemenea să fie făcută înainte ca informațiile să fie colectate. Nu ar fi etic să schimbăm nivelul de încredere după aceea, deoarece deciziile ar fi interpretate ca un efort manipulator pentru a întoarce rezultatele în favoarea sprijinirii concluziilor cercetării.

8.5. Relații statistic semnificative şi rezultate substanțiale

Cuvântul *semnificativ* este larg utilizat si totodată foarte vag în profesia noastră. Îl folosim deseori pentru a accentua importanța a ceva, ca de exemplu "contribuția semnificativă" a unui asistent social la autorizarea plății notelor de plată sau pe rol de "evoluții semnificative" în dezvoltarea respectului de sine la clienți. Cu alte cuvinte îl folosim zilnic, ca o valoare, de aceea este bine să aşezăm alături de înțelesul lui *semnificativ* utilizat în mod obişnuit, şi ceea ce el înseamnă în statistică. ***Semnificația statistică*** **este o demonstrație, prin procedeul testării statistice, care elimină ipoteza nulă într-un mod sigur, şi arată că există o legătură reală între variabile.** O legătură între două variabile care este declarată a fi statistic semnificativă este atunci când suntem în mare măsură siguri (95%, în cele mai multe cazuri) că ea nu poate fi datorată şansei sau erorii de eşantionare. În statistică există doar un singur înțeles relevant al cuvântului semnificativ sau semnificație; trebuie să fim atenți în folosirea termenilor numai în acest sens.

O legătură statistic semnificativă între variabile poate sau nu sugera un rezultat într-adevăr important al cercetării. Trebuie să fim atenți la evaluarea fiecărei legaturi cu suport statistic, în contextul întrebării, "Ei şi ce ?!" În practica de asistență socială, **nu orice legătură statistic semnificativă este un rezultat ce necesită o implementare reală**. In realitate pot exista legaturi statistic semnificative care sunt judecate ca fiind nesemnificative în sens absolut.

Un exemplu ne poate ajuta să ilustrăm această distincție. Un asistent social care a condus un studiu de cercetare pentru a determina care tip de abordare (A sau B) produce rezultate mai bune la clienții care apelează la serviciile de consiliere familială. El demonstrează că scorul mediu de 53 puncte obținut de cuplurile care au primit Tratamentul A este suficient de diferit față de scorul mediu de 57 a cuplurilor ce au primit Tratamentul B. Asistentul social se simte sigur în ceeace priveşte eliminarea ipotezei nule şi concluzionează că există o legătură statistic semnificativă între tipul de tratament şi aranjamentul marital. Dar după o privire mai atentă, el concluzionează că rezultatele nu sunt relevante, pentru că o diferență de numai patru puncte (57-53) este prea mică. Această diferență nu este în mod sigur îndeajuns de mare pentru a justifica trimiterea câtorva membrii ai personalului la un program costisitor de învățare şi deprindere a Tratamentului B. Bazându-se pe lipsa cunoaşterii a ceea ce poate fi interpretat ca un rezultat substanțial (o diferență banală de patru puncte) supervizorul decide să nu implementeze rezultatele.

Existența unei legături statistic semnificative între variabile poate fi determinată prin testare statistică pe baza legilor probabilității. După aceea trebuie determinat dacă un rezultat este suficient de substanțial sau nu. Această decizie cere o bună înțelegere a mai multor aspecte diferite din practica asistenței sociale.

Întrebări pentru studiu

1. Înainte de a pretinde că există o legătură reală între variabile, care sunt cele trei explicații concurente care trebuie eliminate ?
2. Care dintre explicațiile concurente necesită aportul statisticii ca să fie respinse ?
3. Care explicații concurente trebuie să fie controlate la începutul planului de cercetare?
4. Care sunt ceilalți patru termeni pentru şansă folosiți în domeniul testării ipotezelor?
5. Care este diferența dintre o eroare de tipul I şi una de tipul II ?

6. Care este forma nulă a unei afirmații pentru o legătură între vârstă si preferința politică?
7. Care este legătura dintre ipoteza nulă şi şansă în testarea ipotezei ?
8. In ce condiții se folosesc alte nivele de semnificație decât convenționalul 0,05 ?

Exercitiu pentru acasă 7

Care nivel de semnificație 0,01 sau 0,10 sugerează o probabilitate mai mare de existență a unei legături reale între variabile ?

Modul 9

Asocierea

9.1. Ce caută să determine asocierea

Aşa cum am discutat anterior, toate testele statistice încearcă să elimine şansa ca explicaţie pentru o legătură aparentă între două sau mai multe variabile. Cu asocierea, spectrul şansei are anumite caracteristici. Dacă spre exemplu, am căuta să găsim o legătură între o variabilă dependentă ca succesul tratamentului clientului (succes/eşec) şi o variabilă independentă asemenea tipului de tratament (de grup/individual) pentru urmărirea evaluării unui program de consiliere în privinţa alcoolismului, şansa ar putea juca rolul scepticului. Ipoteza nulă spune că nu există nici o legătură între cele două variabile: dacă clienţii s-au abţinut de la alcool pentru o perioadă de timp (succes) sau nu s-au abţinut (eşec) şi dacă au primit tratamentul de grup sau individual. Ocazional poate apărea că acei clienţi care s-au abţinut au fost în general cei care au primit tratamentul individual sau viceversa. Totuşi, în concordanţă cu ipoteza nulă, această aparentă legătură dintre cele dintre două variabile poate fi explicată şi de variaţiile normale ale caracteristicilor întâlnite în eşantionul mic pe care s-a lucrat (faţă de populaţia tuturor alcoolicilor). Ar putea fi doar rezultatul erorii de eşantionare. Ipoteza nulă trebuie să susţină că nu exista o legătură reală între două variabile dintr-o populaţie.

Putem pretinde existenţa unei legături reale între două variabile doar dacă putem demonstra că legătura observată dintre ele este improbabil să se fi datorat şansei, şi deci orice persoană rezonabilă va elimina întâmplarea (şi fireşte factorii de deformare şi alte variabile) ca explicaţie improbabilă.

Asocierea încearcă să determine dacă există o legătură adevărată între două variabile, examinând în ce măsură valorile specifice unei variabile sunt asociate cu valorile specifice ale celei de a doua variabile, într-un grad de probabilitate suficient de mare ca aceasta să nu fie doar efectul erorii de eşantionare.

Când folosim asocierea nu suntem atât de ambiţioşi ca să sugerăm că o variabilă poate fi cauza variaţiei celeilalte variabile. Putem doar afirma, în cel mai bun caz, că există un anumit tip de legătură (patern). Testele de asociere sunt folosite când dorim să ştim dacă aceste legături sunt suficient de puternice şi consistente pentru a elimina şansa (întâmplarea) ca o explicaţie a legăturii observate.

Analiza asocierii este una dintre cele mai simple căi pentru a determina dacă există o legătură adevărată între două variabile. Exemplul care urmează ilustrează modul în care asocierea este folosită în examinarea eficacităţii relative a tratamentului de grup faţă de tratamentul individual la clienţii care au fost trataţi pentru alcoolism într-un centru nonguvernamental de tratament. Pentru a face aceasta, un grup de clienţi care au primit tratamentul în grup sunt comparaţi direct cu un grup de clienţi care au primit tratamentul individual.

9.2. Tabele de asociere

Pentru a înţelege procedeul asocierii este util să începem cu un tabel care este asemănător Tabelului 9.1 sau Tabelului 9.2 . Tabelele au multe nume: ***tabele de asociere, tabele hi-pătrat sau tabele de contingenţă***. În continuare ne vom referi la ele ca tabele de asociere. În ultima coloană din dreapta Tabelului 9.1 şi Tabelului 9.2, sunt totalurile (frecvenţele) pentru fiecare rând introdus. Totalurile pe coloane sunt introduse în linia de jos. Aceste totaluri de pe rânduri şi coloane sunt denumite ***frecvenţe marginale***. Ele indică numărul total de cazuri care au fost observate având o anumită valoare pentru una din variabilele - aceste sunt: tratament în grup, tratament individual, succes sau eşec. Totalul general, ***numărul total de cazuri*** *(N),* este introdus în colţul din dreapta-jos. Suma totalurilor de pe ultima coloana şi de pe ultimul rând sunt egale, şi egale cu numărul total de cazuri.

Aşa cum se vede, Tabelul 9.1 conţine două variabile dihotomice (cu două categorii), tipul de tratament şi succesul clientului. Clienţii diferă după tipul de tratament pe care l-au primit (variabila independentă) şi pot varia datorită succesului (variabila dependentă). Este, desigur, posibil a avea variabile cu mai mult de două categorii. Tabelul de asociere va avea atunci mai multe rânduri şi coloane şi, firesc,

mai multe celule. Frecvențele din interiorul tabelului de asociere, notate in Tabelul 9.1 cu *a, b, c* şi *d*, poartă numele de ***frecvențe celulare***.

Tabelul 9.1 Tabel de asociere pentru succesul in abtinerea de la alcool si tipul de tratament

	Succes?		
Tip de tratament	Da	Nu	Total
Tratament în grup	a	b	a+b
Tratament individual	c	d	c+d
Total	a+c	b+d	N

Categoriile variabilelor din tabelul de asociere pot fi puse în orice ordine, deoarece ele sunt de nivel nominal, neexistând o ordonare după rang sau alte diferențe cantitative. Matematic, este posibil să folosim analiza tabelelor de asociere cu variabile de nivel ordinar, interval sau raport. Totuşi, folosind tabelele de asociere cu variabile de nivel ordinal sau interval, acestea nu vor putea profita de avantajele pe care le oferă precizia de măsurare a acestor variabile. Valorile, în aceste cazuri, sunt tratate fără să se țină cont de diferențele cantitative pe care le reflectă, ca şi cum ele ar reprezenta doar diferențe calitative.

Tabelele de asociere prezintă de obicei frecvențele pentru o variabilă independentă şi pentru una dependentă. În acest capitol, în toate tabelele, variabila dependentă va fi dispusă pe coloane, iar variabila independentă va fi dispusă pe linii. Aceasta nu reprezintă o lege, de aceea unele studii folosesc aşezarea inversă. De fapt, în orice studiu de cercetare, în momentul calculului, indicatorul de asociere este "orb" în ceea ce priveşte care dintre variabile este cea independentă şi cea dependentă. Asocierea examinează numai dacă există legături între cele două variabile. Sunt situații când nici una dintre variabile nu este clar dependentă sau independentă. Ele sunt doar două variabile, a căror legătură dorim să o studiem. Indicatorul hi-pătrat este, de asemenea, potrivit pentru acest tip de situații.

Frecvențe observate

Tabelul 9.2 prezintă rezultatele actuale sau observate, rezultate din studiul nostru ipotetic asupra celor două metode de tratament. Datele din cele patru celule ale Tabelului 9.2 reprezintă numărul observat de clienți care au realizat fiecare combinație de valori corespunzătoare pentru cele două variabile. Putem observa că au existat în total 100 de clienți (N), dintre care 60 au primit tratamentul în grup (a+b) şi 40 au primit tratamentul individual (c+d). Cincizeci şi cinci au avut succes, în timp ce 45 au avut insucces. În plus, printre cei 60 care au primit tratamentul în grup, 40 de clienți au avut succes (celula a) şi 20 de clienți au avut insucces (celula b). Printre clienții care au primit tratamentul individual, 15 au fost considerați ca având succes (celula c) şi 25 ca având insucces (celula d).

Tabel 9.2 Numărul de succese observate, după tipul de tratament

	Succes?		
Tip de tratament	Da	Nu	Total
Tratament în grup	40	20	60
Tratament individual	15	25	40
Total	55	45	100

În exemplul nostru avem nevoie să comparăm clienții care au primit tratamentul în grup cu aceia care au primit tratamentul individual, în ceea ce priveşte rezultatele lor. O astfel de comparație este relativ greu de observat în Tabelul 9.2 pentru că cele două tipuri de tratament au numere diferite de clienți (60 şi 40). Evident că pe cei 40 de clienți care au avut ca rezultat succesul cu tratamentul în grup (celula a), nu-i putem compara direct cu cei 15 clienți care au avut ca rezultat succesul, dar pe baza tratamentului individual (celula c) şi astfel nu putem concluziona că tratamentul în grup este cea mai bună metodă de tratament doar pentru că numărul 40 este mai mare decât 15. În ciuda diferenței dintre numărul de cazuri în cele două grupuri, este posibil să facem o încercare de comparare între cele două tipuri de tratament, prin calcularea procentelor. De exemplu, putem afla ce procentaj reprezintă 40 de clienți din 60 de clienți şi ce procentaj reprezintă 15 clienți din 40 de clienți. Tabelul 9.3 este un tabel de asociere cu procentaje pentru datele observate în Tabelul 9.2. El arată că 66,7 procente ale clienților care au primit tratamentul în grup au avut ca rezultat succesul, comparativ cu 37,5 procente ale acelor clienți care au primit tratamentul

individual. Variabilele tipul de tratament primit şi succesul clienţilor *ar putea* fi deci legate. Dacă procentajele (celula a şi celula c) ar fi identice, cele două variabile, în mod sigur, nu ar fi legate. Până la acest punct, nu putem însă exclude întâmplarea ca explicaţie a legăturii aparente dintre variabile.

Deşi cele două variabile par să fie întrucâtva legate, totuşi s-ar putea spune că ele "nu sunt prea mult legate". Argumentul ar consta în faptul că 29,2 procente diferenţă (66,7 procente - 37,5 procente = 29,2 procente) nu este foarte mult şi s-ar putea ca faptul de a fi legate este doar un rezultat al erorii de eşantionare. Aşa să fie? Multe dintre raţionamentele statistice sunt preocupate să ne ajute să decidem *cât de mare* trebuie să fie diferenţa pentru a putea elimina şansa ca explicaţie posibilă a unei legături aparente între variabile.

Tabel 9.3 Procentul succeselor observate, după tipul de tratament (din tabelul 9.2)

	Succes?		
Tip de tratament	Da	Nu	Total
Tratament în grup	66.7	33.3	100.0
Tratament individual	37.5	62.5	100.0

Tratament în grup	=	Celula a: 40 / 60 = 66.7%
		Celula b: 20 / 60 = 33.3%
		100%
Tratament individual	=	Celula c: 15 / 40 = 37.5%
	=	Celula d: 25 / 40 = 62.5%
		100%

Frecvenţele aşteptate

Cât de mare ar trebui să fie diferenţa între procente pentru ca şansa să fie o explicaţie improbabilă? Putem răspunde la această întrebare prin concentrarea asupra a cât de mult diferă frecvenţele observate faţă de acele frecvenţe pe care noi ne aşteptăm să le găsim mai frecvent, dacă ipoteza nulă ar fi adevărată - acestea sunt *frecvenţele aşteptate.*

Să ne întoarcem la Tabelul 9.2 şi să ne concentrăm doar asupra frecvenţelor marginale. Din totalul de 100 de clienţi, 55 sau 55 % au avut ca rezultat succesul. Dacă tipul de tratament nu este legat de succesul clientului, ar trebui să ne aşteptăm ca aproximativ 55 de procente din totalul clienţilor să aibă succes, *indiferent* de tipul de tratament aplicat. Deşi rezultatele dintr-un eşantion particular nu vor ieşi *exact* în acest mod, foarte adesea, într-un număr mare de eşantioane dintr-o populaţie în care ipoteza nulă este adevărată (variabilele în mod sigur sunt nelegate), vom găsi "rezultatul mediu" al tuturor eşantioanelor; adică proporţia medie de aproximativ de 55 %.

Putem construi un tabel al frecvenţelor aşteptate presupunând că ipoteza nulă ar fi adevărată, asemănător Tabelului 9.4 . Pentru a calcula frecvenţa aşteptată dintr-o celulă, se ia totalul pe coloană, se înmulţeşte cu totalul de pe linie pentru acea celulă şi apoi se împarte rezultatul la numărul total de cazuri (*n*). Adică:

$$f_e = \frac{l * c}{n}$$

unde:

f_e = Frecvenţa aşteptată într-o celulă particulară
l = Totalul pe linia celulei
c = Totalul pe coloana celulei
n = Numărul total de cazuri

Înlocuind valorile vom găsi:

celula a: $f_e = \frac{(60)\,(55)}{100} = 33$

celula b: $f_e = \frac{(60)(45)}{100} = 27$

celula c: $f_e = \frac{(40)(55)}{100} = 22$

celula d: $f_e = \frac{(40)(45)}{100} = 18$

Totalul frecvențelor aşteptate = 100

Tabelul 9.4. Numărul şi procentul succeselor aşteptate, după tipul de tratament

Tip de tratament	Succes? Da		Nu		Total	
	Număr	Procent	Număr	Procent	Număr	Procent
Tratament în grup	33	55%	27	45%	50	100%
Tratament individual	22	55%	18	45%	150	100%
Total	55		45		100	

Tabelul 9.5. Diferența dintre frecvențele observate şi cele aşteptate.

Număr de succese, după tipul tratamentului (din Tabelele 9.2 şi 9.4)

Celule	Observate (Tabelul 9.2)		Aşteptate -(Tabelul 9.4)	=	Diferențe = (Tabelul 9.5)
Celula a	40	-	33	=	+7
Celula b	20	-	27	=	-7
Celula c	15	-	22	=	-7
Celula d	25	-	18	=	+7
Total	100	-	100	=	0

Tip de tratament	Succes? Da	Nu	Total
Tratament în grup	+7	-7	0
Tratament individual	-7	+7	0
Total	0	0	0

Diferența dintre frecvențele observate şi frecvențele aşteptate

Tabelul pentru frecvențele observate este acum comparat cu tabelul frecvențelor aşteptate. Aceasta înseamnă că vom examina mai îndeaproape diferențele dintre frecvențele observate (Tabelul 9.3) şi frecvențele aşteptate (Tabelul 9.4) pentru fiecare celulă. Tabelul 9.5 prezintă diferențele matematice între frecvențele observate şi frecvențele aşteptate din exemplul nostru.

Indicatorul statistic hi-pătrat

Ar fi nevoie acum de un indicator unic al diferențelor prezentate în Tabelul 9.5. Simpla adunare a diferențelor pentru toate celulele nu este utilă pentru că va fi *întotdeauna* zero. Un indicator mai bun poate rezulta prin ridicarea la pătrat a diferențelor din fiecare celulă, împărțirea acestor pătrate la valoarea aşteptată pentru fiecare celulă şi adunarea rezultatelor pentru toate celulele. Numărul care rezultă este numit *valoarea hi-pătrat*, reprezentat de litera din alfabetul grec χ *(hi)* cu semnul ridicării la pătrat. Putem exprima aceasta cu formula:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

A

unde: $\chi 2$= Valoarea hi-pătrat

f_o = Frecvența observată

f_e = Frecvența aşteptată

$\sum$ = Suma (tuturor celulelor)

Înlocuind literele cu valori găsim:

$$\chi^2 = \frac{(40-33)^2}{33} + \frac{(20-27)^2}{27} + \frac{(15-22)^2}{22} + \frac{(25-18)^2}{18} = \ldots = 8{,}2$$

$$\ldots = \frac{(+7)^2}{33} + \frac{(-7)^2}{27} + \frac{(-7)^2}{22} + \frac{(+7)^2}{18}$$

$$= (49)/33 + (49)/27 + (49)/22 + (49)/18$$

$$= 1{,}5 + 1{,}8 + 2{,}2 + 2{,}7$$

$$= 8{,}2 \text{ (valoarea lui hi-pătrat)}$$

Valorile indicatorului statistic χ^2 al tuturor eşantioanelor posibile se distribuie după o curbă matematică bine cunoscută, cu următoarele proprietăți: i) este pozitiv asimetrică, ii) are originea în 0 (zero), din cauza ridicării la pătrat şi iii) e o familie de curbe a căror formă depinde de mărimea tabelului de asociere.

Observație. Dacă tabelul de asociere are doar patru celule (2x2) aşa cum este cazul în exemplul nostru, trebuie să mai scădem 0,5 din diferența dintre frecvențele observate şi cele aşteptate pentru fiecare celulă înainte de ridicarea la pătrat (acesta este Factorul de corecție al lui Yates – in SPSS *Continuity Corection*). Dar de dragul simplificării şi pentru a prezenta formula obişnuită, exemplul nostru nu a mai efectuat această scădere.

Grade de libertate

Înainte să putem utiliza un tabel de asociere pentru calculul lui hi-pătrat, pentru a determina dacă există o asociere statistic semnificativă între cele două variabile, avem nevoie să înțelegem conceptul de grade de libertate. Probabilitatea obținerii unei valori hi-pătrat mari este afectată de mărimea tabelului de asociere pe baza căruia este calculat. Mărimea se referă aici la numărul de coloane şi de linii (adică numărul total de celule) din tabel. Cu cât tabelul este mai mare, cu atât este mai probabil să avem o valoare mai mare al lui hi-pătrat. Aceasta reiese evident din faptul că valoarea hi-pătrat este suma cifrelor derivate din fiecare dintre celule. Cu cât sunt mai multe celule într-un tabel, cu atât vor fi mai multe cifre care adunate, vor creşte valoarea lui hi-pătrat.

Fiecare valoare hi-pătrat trebuie să fie evaluată ținând cont de dimensiunea tabelului, exprimată în termeni de grade de libertate. Numărul de grade de libertate pentru un tabel de asociere este egal cu numărul de linii minus unu, înmulțit cu numărul de coloanelor minus unu. Putem scrie această formulă astfel:

$df = (l\text{-}1)\,(c\text{-}1)$

unde:

df = grade de libertate

l = numărul de linii

c = numărul de coloane

Înlocuind literele cu valorile din exemplul nostru, găsim:

$df = (2\text{-}1)\,(2\text{-}1)$

$= (1)\quad (1)$

$= 1$ *(grade de libertate)*

Într-adevăr, Tabelul 9.2 are un singur grad de libertate, aşa cum au toate tabelele formate din două linii şi două coloane.

Gradele de libertate exprimă câte frecvențe celulare pot fi schimbate în mod liber păstrându-se neschimbate toate frecvențele marginale.

Determinarea probabilității

Pentru a determina dacă valoarea lui hi-pătrat pentru un tabel de asociere dat sugerează sau nu o asociere statistic semnificativă între variabile, trebuie să găsim în primul rând linia care corespunde gradelor de libertate al tabelului de asociere în Tabelul 9.6. Cele şase valori din fiecare linie a Tabelului 9.6 sunt valori hi-pătrat care au probabilitatea indicată în capul de tabel al coloanelor respective. Vom citi în dreptul liniei pentru a găsi unde cade valoarea noastră hi-pătrat. Dacă numărul exact nu apare, vom considera numărul din *stânga* locului unde ar cădea valoarea lui hi-pătrat. După aceea ne vom deplasa la vârful coloanei şi vom găsi probabilitatea asociată lui.

Dacă, de exemplu, fixăm nivelul de probabilitate *P* la 0,95 - ceea ce corespunde unui nivel de semnificație *p* de 0,05 - vom şti că dacă respingem ipoteza nulă, probabilitatea statistică de a comite o eroare de tipul I este mai mică decât 5 din 100.

În exemplul nostru, valoarea obținută pentru hi-pătrat este de 8,2, cu un grad de libertate. Luăm valoarea lui hi-pătrat de 8,2 şi găsim cele două valori din prima linie a Tabelului 9.6 între care se găseşte această valoare. Valoarea noastră hi-pătrat, 8,2 este localizată între valorile 6,64 şi 10,83. Astfel, dacă ipoteza noastră a fost direcțională, adică "clienții care primesc tratamentul în grup au o rată statistic semnificativă mai înaltă de succes decât clienții care primesc tratamentul individual", putem spune că dacă respingem ipoteza nulă, există o probabilitate de doar 0,005 de a face o eroare de tipul I (doar 5 dintr-o mie). Pe scurt, ipoteza noastră direcțională poate fi considerată ca având suport statistic, deoarece 0,005 este mult mai mic decât convenționalul 0,05. Pe de altă parte, dacă ipoteza noastră ar fi fost nedirecțională, putem încă considera că avem suport statistic pentru ea, pentru că probabilitatea corespunzătoare este tot mai mică decât 0,01, care este mai mică decât convenționalul 0,05 .

Tabelul 9.6 Valori critice pentru hi-pătrat

	Nivel de semnificație pentru un test direcțional					
	.10	.05	.025	.01	.005	.0005
	Nivel de semnificație pentru un test nedirecțional					
df	.20	.10	.05	.02	.01	.001
1	1.64	2.71	3.84	5.41	6.64	10.83
2	3.22	4.60	5.99	7.82	9.21	13.82
3	4.64	6.25	7.82	9.84	11.34	16.27
4	5.99	7.78	9.49	11.67	13.28	18.46
5	7.29	9.24	11.07	13.39	15.09	20.52
6	8.56	10.64	12.59	15.03	16.81	22.46
7	9.80	12.02	14.07	16.62	18.48	24.32
8	11.03	13.36	15.51	18.17	20.09	26.12
9	12.24	14.68	16.92	19.68	21.67	27.88
10	13.44	15.99	18.31	21.16	23.21	29.59
11	14.63	17.28	19.68	22.62	24.72	31.26
12	15.81	18.55	21.03	24.05	26.22	32.91
13	16.98	19.81	22.36	25.47	27.69	34.53
14	18.15	21.06	23.68	26.87	29.14	36.12
15	19.31	22.31	25.00	28.26	30.58	37.70
16	20.46	23.54	26.30	29.63	32.00	39.29
17	21.62	24.77	27.59	31.00	33.41	40.75
18	22.76	25.99	28.87	32.35	34.80	42.31
19	23.90	27.20	30.14	33.69	36.19	43.82
20	25.04	28.41	31.41	35.02	37.57	45.32
21	26.17	29.62	32.67	36.34	38.93	46.80
22	27.30	30.81	33.92	37.66	40.29	48.27
23	28.43	32.01	35.17	38.97	41.64	49.73
24	29.55	33.20	36.42	40.27	42.98	51.18
25	30.68	34.38	37.65	41.57	44.31	52.62
26	31.80	35.56	38.88	42.86	45.64	54.05
26	31.80	35.56	38.88	42.86	45.64	54.05
27	32.91	36.74	40.11	44.14	46.96	55.48
28	34.03	37.92	41.34	45.42	48.28	56.89
29	35.14	39.09	42.69	46.69	49.59	58.30
30	36.25	40.26	43.77	47.96	50.89	59.70

32	38.47	42.59	46.19	50.49	53.49	62.49
34	40.68	44.90	48.60	53.00	56.06	65.25
36	42.88	47.21	51.00	55.49	58.62	67.99
38	45.08	49.51	53.38	57.97	61.16	70.70
40	47.27	51.81	55.76	60.44	63.69	73.40
44	51.64	56.37	60.48	65.34	68.71	78.75
48	55.99	60.91	65.17	70.20	73.68	84.04
52	60.33	65.42	69.83	75.02	78.62	89.27
56	64.66	69.92	74.47	79.82	83.51	94.46
60	68.97	74.40	79.08	84.58	88.38	99.61

Valoarea hi^2 este semnificativă dacă ea este mai mare sau egală cu valoarea listată în tabel

Să ținem minte că trebuie să folosim valoarea din *stânga* valorii calculate a lui hi-pătrat pentru a determina corect nivelul probabilității. De exemplu, avem nevoie să găsim o valoare hi-pătrat la cel puțin 2,71, cu un grad de libertate, pentru ca o ipoteză direcțională să fie susținută la un nivel obişnuit de semnificație de 0,05.

Prezentarea rezultatelor unei analize de asociere

Prezentarea rezultatelor noastre obținute în urma unei analize de asociere, este relativ simplă. În primul rând, vom prezenta tabelul de asociere cu frecvențele absolute observate (atât cele absolute, cât şi cele procentuale). Frecvențele observate procentuale trebuiesc calculate prin raportare la frecvențele marginale ale variabilei considerate independente. După aceea plasăm la sfârşitul tabelului valoarea hi-pătrat (χ^2), gradele de libertate (*df*), nivelul de semnificație (*p*) asociat valorii calculate a lui hi-pătrat, şi poziția față de nivelul critic (0.05, sau 0. 01, sau 0.10, după caz). In exemplul nostru, aceste trei elemente de informație vor fi scrise astfel:

$\chi 2 = 8{,}2;\ df = 1\ ;\ p < 0{,}005$

Tabelele 9.7; 9.8; 9.11 şi 9.12 sunt exemple de prezentare a analizei de asociere.

Observații

In SPSS in loc de notația *p* pentru nivelul de semnificație se foloseşte *Sig.*

Pe lângă valoarea χ^2 SPSS afişează in tabelele rezultat şi alți indicatori derivați din acesta -cum ar fi *Coeficientul* φ (phi) sau Coeficientul *V al lui Cramer* - care prezintă avantajul că au limite fixe de variație (între 0 şi 1), ceea ce ne permite să îi interpretăm şi în termeni de indicatori de intensitate ai asocierii (vezi şi Rotariu et. al 1999, p.127 şi p.135). În formula coeficientului *V, s** este minimul valorii dintre numărul de linii şi numărul de coloane. Coeficientul φ se utilizează pentru tabelele 2x2 , iar *V* pentru tabelele de dimensiuni mai mari.

$$\varphi = \sqrt{\frac{\chi^2}{n}} \qquad \text{şi} \qquad V = \sqrt{\frac{\chi^2}{n(s^* - 1)}}$$

9.3. Când să nu utilizăm asocierea

Analiza hi-pătrat poate să ne fie foarte folositoare. Totuşi, ea poate fi greşit înțeleasă. Probabilitatea indicată poate să nu fie prea adecvată în cazul în care frecvențele aşteptate din câteva celule ale tabelului de asociere sunt mici. Există trei situații în care procedeul tabelelor de asociere nu poate fi folosit:

1. Când într-un tabel cu două linii şi două coloane (patru celule), una sau mai multe celule au valoarea aşteptată mai scăzută decât 5.
2. Când într-un tabel cu mai mult decât două linii şi două coloane, există mai mult de 20% din celule care au valori aşteptate mai mici decât 5.

3. Când într-un tabel cu mai mult decât două linii şi două coloane, există celule cu frecvenţe observate nule (0).

O verificare rapidă dacă sunt probleme cu valorile aşteptate prea mici, într-un tabel de asociere poate fi realizată prin localizarea celulei cu valoarea aşteptată cea mai mică. Pentru a face aceasta, se localizează linia şi coloana cu cele mai mici totaluri. Celula cu cea mai mică valoare se află la intersecţia liniei şi coloanei localizate. După aceea, valoarea aşteptată a celulei este determinată cu formula (L) (C) / (N). Dacă frecvenţa aşteptată este 5 sau mai mult, este permisă folosirea analizei tabelului de asociere. Dacă ea este mai mică decât 5, poate fi necesar să combinăm anumite celulele între ele (prin grupare), astfel încât criteriul pentru folosirea lui hi-pătrat să poată fi îndeplinit; sau se poate folosi un alt test statistic (vezi ultimul capitol). Desigur, un tabel 2 x 2 (două linii şi două coloane) nu poate fi grupat.

Volumul eşantionului

În general, cu cât avem un eşantion mai mare, cu atât avem mai multe şanse să respingem ipoteza nulă. Cu cât este mai mare dimensiunea eşantionului, cu atât este mai puternic testul. (Acesta este valabil pentru orice test statistic) De fapt, cu un eşantion foarte mare este extrem de probabil ca ipoteza nulă să fie respinsă, chiar dacă diferenţa absolută dintre frecvenţele aşteptate şi cele observate din fiecare celulă este suficient de mică. Când oamenii interpretează un tabel de asociere, sunt adesea induşi în eroare de valoarea lui hi-pătrat şi de nivelul de probabilitate rezultat, mai ales dacă nu urmăresc cu atenţie volumul eşantionului (N). Trebuie întotdeauna să avem în minte că o valoare hi-pătrat şi nivelul de probabilitate sunt legate *direct* de dimensiunea eşantionului pe baza căruia sunt calculate.

Cele prezentate anterior pot părea greu de înţeles, dar un acelaşi tabel de asociere poate prezenta o legătură statistică importantă între două variabile (via indicatorul statistic hi-pătrat) sau - credeţi sau nu - el poate descrie o legătură *slabă*, dar statistic semnificativă. Pe scurt, putem aproape întotdeauna avea o valoare hi-pătrat statistic semnificativă - interesând mai puţin magnitudinea legăturilor dintre două variabile - dacă eşantionul este suficient de mare. Astfel, trebuie întotdeauna să privim la ceea ce înseamnă legătura - nu doar nivelul de semnificaţie statistică (p) al valorii hi-pătrat. Aceasta este legată de discuţia noastră dintr-un capitol anterior în care am făcut distincţie între: (1) legături statistic semnificative între variabile şi (2) rezultate substanţiale.

O continuare a exemplului nostru va clarifica cele afirmate. Să presupunem că într-un alt studiu, 200 de clienţi au primit tratament în cadrul programului de tratare a alcoolicilor. Rezultatul studiului poate fi asemenea celui prezentat în Tabelul 9.7. Aşa cum poate fi observat din acest tabel, p este mai mare decât 0,20 doar dacă direcţia legăturii nu a fost precizată şi mai mare decât 0,10; dacă s-a specificat direcţia în prealabil (vezi Tabelul 9.6). Cu alte cuvinte, noi am putea să nu avem suport statistic suficient la nivelul 0,05 pentru a fi capabili să respingem ipoteza nulă.

Tabelul 9.7 Testul hi^2, numărul şi procentul succeselor observate, după tipul de tratament pentru un eşantion de 200 clienţi

Tip de tratament	Succes? Da		Nu		Total	
	Număr	Procent	Număr	Procent	Număr	Procent
Tratament în grup	30	60.0%	20	40.0%	50	100%
Tratament individual	80	53.3%	70	46.7%	150	100%
Total	110		90		200	

$\chi 2 = 0{,}672$, $df = 1$; $p > 0{,}20$ (fără predicţia direcţiei)

Acum să presupunem că avem nu doar 200 de clienţi, aşa ca în Tabelul 9.7 ci de zece ori mai mulţi - 2000, iar proporţia celor 2000 clienţi în toate celulele este exact aceeaşi ca şi în cazul eşantionului anterior, prezentat în Tabelul 9.7. Rezultatele se găsesc în Tabelul 9.8.

O privire atentă asupra Tabelelor 9.7 şi 9.8 va arăta că frecvenţele observate în ambele tabele sunt absolut proporţionale una faţă de cealaltă, dar diferenţa între valorile fiecărui hi-pătrat şi nivelele de probabilitate este foarte mare. Frecvenţele observate în Tabelul 9.7 nu sunt statistic semnificative, în timp ce frecvenţele observate în Tabelul 9.8 sunt statistic semnificative (la nivelul 0,01 pentru o ipoteză nedirecţională şi la nivelul 0,005 pentru o ipoteză direcţională). Dacă am fi folosit 20000 clienţi valoarea hi-pătrat ar fi fost de 67,2; dacă am fi folosit 200000 clienţi, hi-pătrat ar fi devenit 672 şi aşa mai departe. Şi

totuşi, cele două tabele 9.7 şi 9.8 sunt aproape identice, singurul lucru care le diferenţiază este numărul de cazuri pe care le-am folosit pentru calcularea celor două mărimi hi-pătrat.

Tabelul 9.8 Testul hi^2, numărul şi procentul succeselor observate, după tipul de tratament pentru un eşantion de 2000 de clienţi

Tip de tratament	Succes? Da		Nu		Total	
	Număr	Procent	Număr	Procent	Număr	Procent
Tratament în grup	300	60.0%	200	40.0%	500	100%
Tratament individual	800	53.3%	700	46.7%	1500	100%
Total	1100		900		2000	

$\chi 2 = 6{,}72$, $df = 1$; $p < 0{,}01$ (fără predicţia direcţiei)

9.4. Asocierea cu trei sau mai multe variabile

De obicei ne concentrăm atenţia mai întâi pe legăturile dintre două variabile. Totuşi, trebuie să avem în vedere că o a treia variabilă poate, într-un anumit fel, "explica" legătura aparentă. În exemplul pe care-l vom folosi, începem în primul rând cu legătura dintre cele două variabile: tipul de tratament şi succesul clientului. Este posibil ca o a treia variabilă, nivelul motivaţiei clientului înaintea intrării la tratament, care nu a fost controlată metodologic, să poate explica aparenta legătură între variabila dependentă şi cea îndependentă. Va trebui să o reverificăm pentru a avea o imagine mai bună asupra legăturii adevărate dintre tipul de tratament şi succes. A treia variabilă, motivaţia clientului, se numeşte ***variabilă de control.***

O modalitate de explorare a efectului celei de-a treia variabile este de a împărţi clienţii noştri după categoriile celei de-a treia variabile şi de a examina legătura dintre variabilele principale, controlându-le astfel prin prisma celei de-a treia variabile. În exemplu nostru, vom împărţi eşantionul în două sub-categorii: cu motivaţie înaltă pentru tratament şi cu motivaţie scăzută pentru tratament. Vom construi după aceea două tabele separate, pentru a urmări legătura dintre variabilele noastre, tipul de tratament şi succesul clientului. Rezultatul poate apărea ca în Tabelul 9.9 (motivaţie înaltă pentru tratament) şi Tabelul 9.10 (motivaţie scăzută pentru tratament). Legătura dintre tipul de tratament şi rezultatul clientului aproape că a dispărut, aşa cum putem vedea examinând diferenţele dintre frecvenţele observate şi cele aşteptate în celulele respective (ele sunt aproape zero). Astfel, controlând motivaţia clienţilor pentru tratament, legătura aparentă dintre variabilele dependentă şi independentă nu mai are suport statistic. Este foarte probabil ca legătura observată dintre variabila dependentă şi cea independentă să nu fi fost una reală, ci una *înşelătoare* (*spurious* in limba engleză). Legătura dintre tipul de tratament şi succes a putut apare pentru că motivaţia este asociată cu ambele variabile din ipoteză: cu succesul (pentru că o motivaţie ridicată influenţează direct rata succesului) pe de o parte, şi cu metoda de tratament (pentru că subiecţii înalt motivaţi au perseverat mai mult în a participa la şedinţele de terapie în grup) pe de altă parte.

Tabelul 9.9 Testul hi^2, numărul şi procentul succeselor observate, după tipul de tratament pentru subeşantionul clienţilor cu o motivaţie înaltă (N = 70)

Tip de tratament	Succes? Da		Nu		Total	
	Număr	Procent	Număr	Procent	Număr	Procent
Tratament în grup	22	55.0%	18	45.0%	40	100%
Tratament individual	16	53.3%	14	46.7%	30	100%
Total	38		32		70	

$\chi 2 = 0{,}02$, $df = 1$; $p > 0{,}20$ (pentru un test nedirecţional)

Tabelul 9.10 Testul hi^2, numărul şi procentul succeselor observate, după tipul de tratament pentru subeşantionul clienţilor cu o motivaţie scăzută (N = 30)

Tip de tratament	Succes? Da		Nu		Total	
	Număr	Procent	Număr	Procent	Număr	Procent
Tratament în grup	11	55%	9	45%	20	100%
Tratament individual	6	60%	4	40%	10	100%
Total	17		13		30	

$\chi 2 = 0{,}68$, $df = 1$; $p > 0{,}20$ (pentru un test nedirecţional)

Nu întotdeauna o relaţie iniţial confirmată statistic dispare pentru că ea ar fi înşelătoare. Atunci când cea de a 3-a variabilă introdusă în analiză se situează între variabila independentă şi cea dependentă în sensul că variabla independentă influenţează de fapt variabila de control, iar aceasta la rândul ei este cea care influenţează variabila dependentă, avem de a face cu variabilă *înnăbuşită* (se mai numeşte şi *latentă*). De exmplu am putea controla legătura observată dintre metoda de tratament şi succes după timpul petrecut în şedinte de terapie. N-ar fi de mirare să dispară şi în acest caz suportul statistic pentru ipoteza iniţială, dar explicaţia ar fi alta: metoda de tratament în grup cere participarea periodică la şedinţe făcând astfel să crească timpul acordat terapiei, iar timpul în sine duce la creşterea succesului în abţinerea de la băutură.

Legătura iniţială nu dispare întotdeauna când o controlăm printr-o a treia variabilă. Fireşte, poate rămâne în esenţă aceeaşi în toate categoriile celei de-a treia variabile. În astfel de cazuri, vom putea concluziona că cea de a treia variabilă nu joacă un rol important în explicarea legăturii iniţiale. Legătura poate fi mai scăzută chiar dacă nu dispare. În acest caz, a treia variabilă poate explica doar o parte, nu totul, dintr-o legătura iniţială. Intensitatea legăturii poate să crească când o a treia variabilă este verificată. În asemenea situaţii, cea de-a treia variabilă este tot o variabilă înăbuşită (sau latentă) aceasta ascunzând nodul real al asocierii dintre variabila dependentă şi cea independentă.

Astfel putem da peste un alt rezultat când introducem o a treia variabilă. Legătura dintre primele variabile poate fi diferită pentru diferitele categorii ale variabilei de control. Acesta nu este un rezultat simplu, el este adesea unul important. Nu întotdeauna este posibil să obţinem uşor rezultate rezumative; mai curând legătura iniţială trebuie să fie descrisă pentru fiecare categorie a variabilei de control. Cea de a treia variabilă este introdusa ca să detaileze mai departe legătura dintre primele două variabile, şi se numeşte variabilă de control.

9.5. Un exemplu

Descrierea situaţiei care generează un studiu

Ca asistent social într-un spital, Ioana se ocupă de internarea pacienţilor. Ea a observat că un număr mare de pacienţi care au fost lăsaţi să trăiască cu propriile rude sunt reinternaţi în spital. Cunoscând că asistenţii sociali, colegi de ai ei, care se ocupă cu planificarea externărilor trimit frecvent pacienţii externaţi la internat, ea s-a întrebat de ce a văzut atât de puţine reinternări printre acei pacienţi care au fost externaţi la internat. Ea s-a întrebat dacă nu poate fi o legătură între pacienţii care sunt reinternaţi în spital şi locul în care au fost ei externaţi (internat/la rude).

Ipoteza ce urmează a fi testată

Ioana citeşte literatura de specialitate asupra temei care o preocupă. Bazându-se apoi pe consensul general al altor practicieni de asistenţă socială, pe rezultatele cercetărilor anterioare şi pe propriile intuiţii şi observaţii subiective, ea porneşte la realizarea şi implementarea unei cercetări de mici proporţii care va strânge datele necesare testării unei ipoteze direcţionale.

Pacienţii externaţi la internat vor avea o rată de reinternare mai scăzută faţă de pacienţii externaţi la rude, statistic semnificativă.

O privire asupra metodologiei

Ioana a ales o strategie simplă pentru a testa ipoteza sa direcţională. Ea a primit permisiunea supervizorilor săi să selecteze un eşantion de 10 procente din toate dosarele pacienţilor care au fost externaţi în ultimele 18 luni, alese la întâmplare. Utilizând un instrument de colectare a informaţiilor standardizat întocmit de ea, a strâns date de o mare varietate a variabilelor demografice pentru 148 de

pacienți (10% din 1480 pacienți = 148 pacienți) care au fost externați la internat şi 250 de pacienți (10% din 2500 pacienți = 250 pacienți) care au fost trimişi la rude. Eşantionul total a fost de 398 pacienți (148+250=398). Variabila dependentă în ipoteza ei a fost dacă pacientul a fost reinternat spital (da/nu). Variabila independentă a fost locul externări pacienților (internat/rude).

Rezultatele

Tabelul 9.11 prezintă rezultatele la care a ajuns Ioana, folosind procedeul tabelelor de asociere aşa cum sunt prezentate în acest capitol.

Tabelul 9.11 Reinternarea în spital după starea externării

	Reinternare?		
Locul externării	Da	Nu	Total
La internat	25	123	148
La rude	71	179	250
Total	96	302	398

$\chi 2 = 7{,}2$, $df = 1$, $p < 0.005$ (utilizând corecția lui Yates)

Interpretarea rezultatelor şi tragerea concluziilor

Ce a aflat Ioana din testarea ipotezei direcționale folosind tabelul de asociere? Din cunoştințele sale generale despre testarea ipotezelor, ea ştia că $p < 0.005$, este un nivel de probabilitate impresionant. Aceasta a însemnat pentru ea că diferențele dintre frecvențele observate şi cele aşteptate au fost mari. Ea ştia, de asemenea, că dacă respinge ipoteza nulă pe baza analizei sale, va greşi de mai puține ori decât 5 dintr-o mie. Astfel, ea a putut respinge ipoteza nulă şi să concluzioneze că există o legătură statistic semnificativă între cele două variabile. Important este că ea a avut suport statistic pentru ipoteza ei direcțională.

De asemenea, Ioana ştia că în analiza tabelei de asociere trebuie să privească nu numai dacă rezultatul este statistic semnificativ, dar şi dacă legătura între cele două variabile a fost în direcția ipotezei. Asemenea altor teste statistice despre care vom discuta, tabelul de asociere nu ia în seamă direcția ipotezei. Deoarece în analiza tabelului de asociere se ține cont în primul rând de *diferențele* dintre frecvențele aşteptate şi cele observate pentru toate celulele, el va fi sensibil la relativa mărime sau micime a frecvențelor observate pentru fiecare celulă, neținând cont de ceea ce a fost prevăzut. Mai trebuie să ne reamintim că o diferență este doar o diferență, chiar dacă sugerează numere mai mici sau chiar mai mari decât cele prezise. O diferență mare între frecvențele observate şi cele aşteptate dintr-o celulă (*în orice direcție*) contribuie mult la creşterea valorii lui hi-pătrat, care se va reflecta prin creşterea probabilității ca ipoteza nulă să fie respinsă. Trebuie să determinăm dacă asocierea este în direcția prezisă, privind direct celulele în care se găsesc frecvențele observate relativ mari, sau examinând procentajele.

Folosind Tabelul 9.11, Ioana a fost capabilă să determine că aproximativ 17% (25 din 148) din pacienții externați la internat au fost readmişi în spital, comparativ cu 28% (71 din 250) dintre aceia dirijați către rude. Aceste două procentaje, 17 şi 28 au fost consecvente cu direcția ipotezei sale; pacienții eliberați către internat au fost mai puțin întâlniți ca reinternați față de pacienții externați la rude.

Înainte ca Ioana să tragă orice concluzie despre "însemnătatea" semnificației statistice dintre cele două variabile, ea a ştiut că trebuie să recunoască efectele metodologiei de cercetare pe care a folosit-o în interpretarea rezultatelor obținute. Ea a folosit un instrument de colectare a datelor standardizat şi structurat. Totuşi, validitatea şi siguranța informațiilor din fişele pacienților poate fi o problemă, ca şi alți factori deformatori. Din cauza lipsei unui plan experimental, lista altor variabile (factori) care ar fi putut afecta reinternarea ar putea fi mare. Printre aceştia s-ar putea număra: diagnosticul pacientului, durata primei spitalizări, disponibilitatea serviciului de îngrijire de după externare, medicația folosită de pacient şi mulți alți factori pe care ea nu are motiv să creadă că au fost egal reprezentați în cele două grupuri de pacienți (îngrijiți acasă / în internat).

Deci, ce îi spun rezultatele despre ipoteză? Scopul procedeului tabelelor de asociere este de a căpăta probe pentru sau împotriva existenței unei legături între două variabile. Cunoaşterea relației cauză-efect nu este posibilă de la început, datorită absenței unui plan experimental şi datorită limitelor proprii ale analizei de asociere. Ceea ce Ioana a aflat este faptul că pentru diferite motive, pacienții externați din spitalul ei către internate, au avut o probabilitate mai mică să fie reinternați față de aceia care au fost eliberați acasă.

Ioana nu şi-a limitat analiza asocierii doar la legătura dintre variabila independentă şi cea dependentă. Ea a mai adunat date despre diagnosticul pacienţilor şi durata primei spitalizări. Ea a putut deci examina legătura dintre aceste "alte variabile" şi variabila dependentă folosind mai multe analize complexe ale tabelelor de asociere, şi alte teste statistice adecvate. Fişele pacienţilor pot conţine informaţii despre variabile suplimentare care au contribuit la luarea deciziei de externare, cum ar fi unde au locuit înainte de internare (la rude sau în internat); aceste informaţii pot fi folosite pentru a tempera rezultatele analizei sale şi pentru a lăsă să cadă mai multă lumină pe rezultatele statistice.

9.6. Folosirea SPSS în analiza asocierii

Pentru a exemplifica asocierea cu ajutorul programului SPSS, vom apela la fişierul de date „1991 US General Social Survey". Vom încerca să vedem dacă există vreo asociere între sexul respondenţilor (variabila *sex*) şi consumul de droguri (variabila *hlth5*). Vom considera sexul ca variabilă independentă şi consumul de droguri ca variabilă dependentă, ambele variabile fiind dihotomice (cu doar două variante de răspuns). Ipoteza de la care pornim este că bărbaţii consumă într-o măsură mai mare droguri decât femeile.

Opţiunile de meniu pe care le vom aplica în SPSS sunt: ***Statistics – Summarize – Crosstabs***. Fereastra de dialog deschisă de *Crosstabs* conţine lista tuturor variabilelor fişierului de date, din care vom selecta variabilele ale căror categorii vor constitui rândurile tabelului (*Rows* – variabila *sex*), respectiv coloanele (*Columns* – variabila *hlth5*). Butonul *Statistics* din aceeaşi fereastră de dialog deschide, la rândul său, o fereastră în care putem selecta coeficienţii de asociere care dorim să fie calculaţi, în cazul nostru hi pătrat - χ^2 (*Chi square*). De la butonul *Cells*, care apare tot în fereastra deschisă de opţiunea de meniu *Crosstabs*, vom deschide o altă fereastră în care vom opta ca în căsuţele tabelului să apară atât valorile observate, cât şi cele calculate pentru cazul independenţei (frecvenţele aşteptate sau teoretice): *Counts – Observed, Expected*. De asemenea, vom opta aici pentru procente pe linii: *Percentages – Row*. Iată ce ne va afişa programul SPSS:

```
CROSSTABS
  /TABLES=sex  BY hlth5
  /FORMAT= AVALUE TABLES
  /STATISTIC=CHISQ
  /CELLS= COUNT EXPECTED ROW .
```

Crosstabs

Respondent's Sex * Illegal Drugs (Marijuana, Cocaine) Crosstabulation

			Illegal Drugs (Marijuana, Cocaine)		
			Yes	No	Total
Respondent's Sex	Male	Count	23	393	416
		Expected Count	12,3	403,7	416,0
		% within R's Sex	5,5%	94,5%	100,0%
	Female	Count	7	589	596
		Expected Count	17,7	578,3	596,0
		% within R's Sex	1,2%	98,8%	100,0%
Total		Count	30	982	1012
		Expected Count	30,0	982,0	1012,0
		% within R's Sex	3,0%	97,0%	100,0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	16,149[b]	1	,000		
Continuity Correction[a]	14,670	1	,000		
Likelihood Ratio	16,190	1	,000		
Fisher's Exact Test				,000	,000
Linear-by-Linear Association	16,133	1	,000		
N of Valid Cases	1012				

a. Computed only for a 2x2 table

b. 0 cells (,0%) have expected count less than 5. The minimum expected count is 12,33.

Observăm că mai întâi programul SPSS afişează sintaxa corespunzătoare opţiunilor de meniu executate. Primul tabel este tabelul de asociere dintre variabila sex şi variabila consum de droguri. Pentru că am cerut calcularea procentelor pentru categoriile variabilei independente, putem observa o diferenţă între ponderea bărbaţilor şi femeilor care au raportat consumul de droguri (5.5% faţă de 1.2%). Ipoteza noastră pare să fie confirmată de aceste date.

Următorul tabel conţine valoarea coeficientului de asociere cerut, împreună cu teste de semnificaţie privind diferenţa lor faţă de zero. Observăm că valoarea coeficientului de corelaţie χ^2 (16,149), ne permite respingerea ipotezei nule, nivelul de încredere fiind mai mare de 99,9% ($p<0.001$). Aşadar, există diferenţe semnificative între bărbaţi şi femei în privinţa consumului de droguri.

Întrebări pentru studiu

1. Ce reprezintă numerele din fiecare celulă a unui tabel de asociere?
2. Ce se pierde când se foloseşte testul hi-pătrat pentru date care sunt de nivel interval şi normal distribuite?
3. Poate indicatorul statistic hi-pătrat să ne spună dacă o variabilă produce variaţia în a doua variabilă? Explicaţi.
4. Ce sunt frecvenţele aşteptate şi cum sunt ele folosite în testul hi-pătrat?
5. Cum intervin gradele de libertate în determinarea semnificaţiei statistice a unei valori hi-pătrat precizate (de exemplu 10)?
6. Care este valoarea minimă a frecvenţei aşteptate necesare pentru folosirea lui hi-pătrat?
7. Care sunt cei doi paşi ai procesului de determinare a suportului statistic pentru o ipoteză direcţională?
8. Cum poate fi folosit hi-pătrat pentru a examina legătura dintre două variabile când se ţine sub control efectului unei a treia variabile?

Exerciţiu pentru acasă 8

Deschideţi fişierul SPSS „1991 U.S. General Social Survey „. Testaţi ipoteza că femeile se declară în general mai puţin fericite decât bărbaţii (variabilele “sex” şi “happy”). Scrieţi un scurt comentariu al rezultatelor obţinute.

Modul 10

Compararea mediilor

Am prezentat până acum un test popular (hi-pătrat) care este folosit frecvent pentru analizarea legăturii dintre două variabile de nivel nominal şi am discutat apoi despre un alt test (coeficientul de corelaţie *r*, a lui Pearson) destinat examinării relaţiei dintre două variabile de nivel interval sau de rapoarte. In acest capitol, vom examina unul dintre testele statistice potrivite atunci când una din variabile, de obicei cea dependentă, este cel puţin de nivel interval iar cealaltă variabilă, de obicei variabila independentă, este de nivel nominal.

10.1. Popularitatea statisticii *t*

Există câteva feluri diferite de teste proiectate pentru diferite necesităţi de analizare a informaţiilor. Vom prezenta doar unul dintre acestea. Testul *t* este foarte popular şi în special foarte folositor pentru cercetarea în asistenţă socială.

Ce îl face pe testul *t* (numit şi testul Student) atât de folositor? Una dintre cele mai mari nevoi pentru asistenţi sociali este în acele studii care caută evaluarea eficacităţii diferitelor abordări ale intervenţiei. Avem nevoie să ştim dacă o metodă de intervenţie este în mod real mai bună decât alta pentru obţinerea rezultatelor dorite. Testul *t* este ideal pentru a fi folosit în situaţii în care avem două categorii a unei variabile independente (intervenţia A / intervenţia B) care este de nivel nominal, pe când variabila dependentă (comportamentul, atitudinea, valoarea sau orice altceva) pe care sperăm să o găsim afectată, este măsurată pe o scală cantitativă (interval sau rapoarte).

De obicei avem măsurători rezonabil de precise ale variabilei dependente. Exemplele pot include variabile ca: stima de sine, aspiraţia către bunăstare, satisfacţia maritală, nivelul de funcţionare socială, sau alte caracteristici ale clientului. Mulţumită muncii cercetătorilor care au îmbunătăţit instrumentele pentru aceste măsurări, acum aceste variabile pot fi tratate statistic, ca fiind măsurate la nivel interval. Alte variabile dependente frecvent folosite ca: numărul întâlnirilor ratate, numărul împotrivirilor exprimate faţă de un program de bunăstare socială, sau numărul comportamentelor violente îndreptate către soţie, sunt de nivel interval sau de rapoarte datorită naturii lor. Dacă putem justifica o variabilă ca fiind de nivel interval sau de rapoarte şi cealaltă variabilă este doar de nivel nominal, *t* poate fi ales ca test statistic.

Alt motiv pentru care cercetătorii în asistenţă socială adesea folosesc testul *t*, este dimensiunea relativ mică a eşantionului necesar pentru utilizarea acestui test. Eşantioanele mari sunt rar disponibile în cercetarea în asistenţă socială, în special în situaţii clinice. Testul *t* este potrivit pentru cercetările care folosesc un număr mic de cazuri (de exemplu, un total de doar 20 sau 30 de cazuri). În plus cele două grupuri nu trebuie neapărat să conţină un număr egal de cazuri. Chiar şi în cercetările planificate cu mare grijă, de exemplu cele asupra eficacităţii unui tratament, există oameni care renunţă înainte ca studiul să fie terminat. Aceasta conduce la o situaţie în care există mai multe cazuri într-un grup decât în altul. Discrepanţa dintre dimensiunea celor două grupuri este automat controlată de formula lui *t*.

10.2. Logica statisticii *t*

Testul *t*, asemenea altor teste statistice, este folosit pentru a ne ajuta să determinăm dacă o legătură aparentă dintre două variabile este o legătură adevărată, la nivelul întregii populaţii, sau este rodul şansei. Cu *t*, se face de fapt o comparare între două medii. Eşantionul este divizat în două grupuri (sub-eşantioane) pe baza valorii fiecărui caz pentru una din cele două categorii, variabila de nivel nominal (de exemplu, intervenţia A sau intervenţia B). Mărimile mediilor variabilei de nivel interval (sau de rapoarte) pentru cele două grupuri sunt comparate apoi folosind formula lui *t*.

Mediile celor două subeşantioane sunt probabil diferite; *t* este o analiză a mărimii acestei diferenţe. Dacă diferenţa se dovedeşte ca fiind mică din punct de vedere matematic, atât de mică încât hazardul ar putea fi explicaţia probabilă pentru această diferenţă, ipoteza nulă nu va putea fi respinsă. Vom putem atunci concluziona că, mult mai probabil, diferenţa dintre mediile celor două grupuri se datorează şansei şi nu reflectă o diferenţă reală care ar exista în populaţie. Dar dacă testul nostru *t*, indică faptul că diferenţa este destul de mare ca ea să nu fie explicată doar prin jocul şansei, vom fi abilitaţi să respingem ipoteza nulă si să concluzionăm că diferenţa observată în eşantion este într-adevăr o reflectare a unei

diferenţe reale în interiorul populaţiei. Şansa, sub forma erorii de eşantionare, va fi eliminată ca explicaţie pentru o legătură observată între două variabile.

Dacă eşantionul este mic, chiar şi o diferenţă mare între două medii poate fi atribuită şansei. Dar există un punct unde diferenţa dintre doua medii este suficient de mare, astfel ca şansa să fie puţin probabil, o explicaţie pentru legătura aparentă dintre două variabile. Unde poate fi găsit acest punct? Testul *t* ni-l spune. Acest punct se determină sub forma unei probabilităţ statistice, probabilitatea de a face o eroare de tipul I, adică de a respinge ipoteza nulă şi de a concluziona că diferenţa dintre cele două medii este legată de prezenţa valorilor diferite pentru variabila secundară (cea de nivel nominal).

Să luăm un exemplu pentru a ilustra utilizarea lui *t*. Un ghid teoretic de studiu a fost realizat pentru a ajuta la pregătirea absolvenţilor pentru examenul de licenţă. Pentru a evalua rezultatele unui asemenea ghid de studiu, putem selecta la întâmplare 15 din 30 de studenţi care au planificat să participe la examen şi cărora le vom furniza o copie a acestui ghid. Le vom da astfel câteva direcţii specifice pentru a-şi împărţi timpul de studiu, fiecare putând folosi ghidul ca îndrumător.

Cei 15 studenţi care au folosit ghidul vor putea fi priviţi ca un grup experimental, rămânând 15 care nu au folosit ghidul şi care vor fi priviţi ca şi grup de control. După ce cei 30 de abolvenţi vor reuşi să promoveze examenul de licenţă, rezultatele lor vor fi comparate. Nu vom compara direct notele tuturor acelora care au folosit ghidul de studiu cu notele celor care nu l-au folosit. În schimb vom compara media notelor rezultate pentru cei 15 absolvenţi din grupul experimental cu media notelor celorlalţi 15 absolvenţi din grupul de control.

Odată cu compararea mărimilor celor două medii, este necesar să ne punem câteva întrebări:

Este oare diferenţa dintre mediile celor două grupuri suficientă pentru a ne permite să respingem ipoteza nulă, care spune că nu există nici o legătură între folosirea respectiv nefolosirea ghidului de studiu şi rezultatele lor la examen?

Cât de încrezători putem fi, pentru a spune că diferenţa nu s-a datorat erorii de eşantionare? Altfel spus, este posibil ca diferenţa dintre mediile celor două grupuri să reflecte o asociere reală între cele două variabile?

Cu ajutorul testului *t* putem compara mărimile celor două medii, folosind o formulă matematică concepută pentru a ne spune dacă diferenţa dintre mărimile mediilor celor două grupuri este suficient de mare pentru ca legătura dintre cele două variabile să fie puţin probabilă din cauza şansei. Testul *t* ne spune dacă ipoteza nulă poate fi respinsă şi dacă există suport statistic pentru a susţine că ghidul de studiu poate în mod real îmbunătăţi rezultatele la examen. Desigur, chiar dacă se demonstrează că există o legătură statistic semnificativă între cele două variabile, va mai trebui să decidem dacă legatura este una substanţială. Diferenţa rezultată din examinare este suficient de mare ca să justificre preţul de achiziţie a ghidului de studiu?

10.3. Calcularea şi interpretarea valorilor lui *t*

Datorită extinderii utilizării calculatoarelor, aflarea lui *t* nu mai este o problemă în ceea ce priveşte calculul matematic. Acest fapt este o uşurare, deoarece formula lui *t* este complexă şi va implica o cantitate considerabilă de timp dacă se lucrează cu creionul şi hârtia. În plus, există mai multe variante de formule de calcul a valorii *t* în funcţie de natura eşantioanelor: dacă avem două eşantioane diferite de subiecţi (Independent Samples t Test) sau în cazul aceloraşi subiecţi la momente diferite de timp (Paired Sample Test). De asemenea exista diferenţe in calculul statisticii t in ipoteza ca cele două eşantioane fac parte din populaţii cu varianţe egale sau dacă cele două populaţii au grade de eterogenitate diferite (varianţe ne-egale). In cele ce urmează abordam doar cazul in care varianţele celor doua populaţii din care provin eşantioanele sunt egale.

Testul t pentru eşantioane (mici) independente – un exemplu

Ana lucrează ca asistent social într-un program de prevenţie privind infecţia cu HIV – SIDA. În calitate de coordonator de program, pregăteşte două echipe de voluntari pentru implementarea programului într-una din şcolile profesionale din oraş. Una dintre echipe va utiliza metode de informare indirectă, prin distribuirea de pliante şi broşuri elevilor, iar cealaltă echipă va folosi metode interactive, respectiv discuţii şi activităţi de grup directe cu clasele de elevi. După implementarea programului, ea evaluează rezultatele activităţilor de prevenţie, încercând să afle dacă una dintre metodele folosite a fost mai eficientă decât

cealaltă, cu alte cuvinte dacă există diferenţe între cunoştinţele elevilor privind riscurile infecţiei cu HIV în funcţie de modalitatea de prevenţie folosită. În acest sens, utilizează un proces de selecţie separat a elevilor în funcţie de metoda de prevenţie utilizată (adică fiecare eşantion în parte este reprezentativ în raport cu variabila „nivel de informare privind riscurile infecţiei cu HIV") şi le aplică un test de cunoştinţe, al cărui scor minim este 0, iar cel maxim 10. Rezultatele obţinute sunt următoarele: pentru primul eşantion: 5, 7, 8, 9, 6, 7, 10, 8, 6; pentru al doilea eşantion: 9, 10, 8, 6, 8, 7, 9. Pentru a răspunde la această întrebare, se aplică testul *t*, calculându-se mai întâi m_1, m_2, s_1 şi s_2 (mediile şi abaterile standard ale rezultatelor obţinute la testul de cunoştinţe aplicat elevilor din cele două eşantioane), iar apoi valoarea lui *t*, după formula:

$$t = \frac{m_1 - m_2}{\sqrt{\frac{(n_1 - 1) * s_1^2 + (n_2 - 1) * s_2^2}{n_1 + n_2 - 2} * \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Cele două medii m_1 şi m_2 sunt:

$$m_1 = \frac{\sum X_1}{n_1} = \frac{5 + 7 + ... + 6}{9} = \frac{66}{9} = 7{,}33$$ pentru grupa tratata cu metode indirecte

$$m_2 = \frac{\sum X_2}{n_2} = \frac{9 + 10 + ... + 9}{7} = \frac{57}{7} = 8{,}14$$ pt. grupa tratată cu metode interactive

Pentru a calcula s_1 şi s_2 (adică abaterile standard in cele doua eşantioane[2]) ne sunt de ajutor următoarele tabele de calcul:

X_1	k_1	$x_1 * k_1$	$x - m_1$	$(x - m_1)^2$	$(x - m_1)^2 * k_1$
5	1	5	-2,33	5,44	5,44
6	2	12	-1,33	1,78	3,56
7	2	14	-0,33	0,11	0,22
8	2	16	0,67	0,44	0,89
9	1	9	1,67	2,78	2,78
10	1	10	2,67	7,11	7,11
	∑= 9	**∑= 66**			**∑= 20**
		m_1 = 7,33			**$s_1^2 = \sum/(n_1 - 1)$ = 2,5**

X_2	k_2	$x_2 * k_2$	$x - m_2$	$(x - m_2)^2$	$(x - m_2)^2 * k_2$
6	1	6	-2,14	4,59	4,59
7	1	7	-1,14	1,31	1,31
8	2	16	-0,14	0,02	0,04
9	2	18	0,86	0,73	1,47
10	1	10	1,86	3,45	3,45
	∑= 7	**∑= 57**			**∑= 10,86**

[2] Spre deosebire de eşantioanele mari unde abaterile standard σ se calculează ca valori medii ale abaterilor pătratice fata de medie – cu alte cuvinte prin împătrirea sumei diferentelor la pătrat la volumul eşantionului *n*, urmată de extragerea radicalului –, la eşantioanele mici se calculează o abatere standard ajustată *s*, unde in loc de *n* se utilizează *n-1* la numitor. De aici şi diferenţa in notaţie: *s* in loc de σ.

$m_2 = 8,14$	$s_2^2 = \sum/(n_2-1) = 1,81$

$$s_1^{\,2} = \frac{\sum (X_1 - m_1)^2 k_1}{n_1 - 1} = \frac{20}{8} = 2,5$$

$$s_2^{\,2} = \frac{\sum (X_2 - m_2)^2 k_2}{n_2 - 1} = \frac{10,86}{6} = 1,81$$

Acum putem calcula numitorul formulei lui *t,* adică eroarea standard comună *Es*:

$$Es = \sqrt{\frac{(n_1 - 1) * s_1^2 + (n_2 - 1) * s_2^2}{n_1 + n_2 - 2} * \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} =$$

$$= \sqrt{\frac{8 * 2,5 + 6 * 1,81}{14} * \left(\frac{1}{9} + \frac{1}{7}\right)} =$$

$$= \sqrt{\frac{20 + 10,92}{14} * (0,25)} = \sqrt{0,560} = 0,748$$

şi

$$t = \frac{7,33 - 8,14}{0,748} = -\frac{0,81}{0,748} = -1,08$$

Pentru determinarea nivelului de probabilitate a valorii *t* obținută se procedează astfel:

- Se calculează gradele de libertate (*df*) după formula
  $df = n_1 + n_2 - 2$
- În Tabelul 10.1 cuprinzând „valorile critice ale lui *t*", se caută linia corespunzătoare gradelor de libertate *df* (afişate pe prima coloană a tabelului).
- Se compară valoarea *t* obținută (valoarea absolută, neținând cont de semnul valorii) cu valorile tabelate fie pentru testul direcționat (one-tailed), fie pentru cel nedirecționat (two-tailed), în funcție de tipul ipotezei de cercetare.

În eşantionul din exemplu, valoarea lui *t* este -1,08, la 14 grade de libertate (*df* = 16 – 2 = 14). Întrebarea este: "Care e probabilitatea de a comite o eroare de tipul I cu o valoare a lui *t* de -1,09, la 14 grade de libertate date?" Pentru a răspunde la această întrebare, avem nevoie să consultăm tabelul de probabilități al valorilor *t* (Tabelul 10.1). Dacă valoarea *t* calcuată este mai mare decât valoarea critică de 2.145 (corespunzător la 14 grade de libertate), nivelul de semnificație *p* va fi mai mic decât 0.05 pentru o ipoteză nedirecționată şi decât 0.025 pentru o ipoteza direcționată, deci se va respinge ipoteza nulă.
În exemplul nostru, valoarea *t* nu este statistic semnificativă, atât pentru un test cu o ipoteză nedirecționată (two-tailed), cât şi pentru un test direcționat (one-tailed), deoarece 1.09 este mai mic decât 2.145.

Interpretare: De vreme ce 1.08 < 2.145, Ana acceptă ipoteza nulă şi concluzionează că nu există diferențe în capacitatea unei metode de prevenție de a duce la cunoştințe mai solide privind riscul infecției cu HIV în rândul elevilor. Rezultatele obținute i-au arătat Anei că nu există suport statistic pentru ipoteza sa, în ciuda faptului că valoarea medie a grupului de elevi cu care s-a lucrat prin metode indirecte indică un

nivel de cunoştinţe mai redus decât media grupului elevilor care au participat la activităţi interactive. Imposibilitatea de a respinge ipoteza nulă relevă că metoda de prevenţie folosită produce mici diferenţe între cunoştinţele elevilor despre riscurile contaminării cu HIV. Firește, Ana şi-a pus întrebarea dacă ipoteza de cercetare (aceea a diferenţelor între cunoştinţele elevilor în funcţie de metoda folosită) poate fi încă corectă. S-ar putea ca adevărata legătură dintre variabila dependentă (cunoştinţele dobândite în urma programului de prevenţie) şi cea independentă (metoda de prevenţie folosită) să fi fost ascunsă de tehnica măsurătorilor sau de influenţa altor variabile (de exemplu, experienţa mai mare a voluntarilor care au aplicat metode interactive). Ana este conştientă că sunt necesare studii suplimentare pentru a clarifica acest aspect. În orice caz, pornind de la rezultatele acestui mini studiu, ea va stabili dacă, pe viitor, va păstra folosirea unor metode diferite de prevenţie de la un grup de elevi la altul sau va merge pe combinarea metodelor de lucru indirecte cu cele directe. Cum rezultatele acestui studiu nu justifică deocamdată realizarea unor schimbări în modul de implementare a programului de prevenţie la elevi, Ana va continua punerea în aplicare a acestuia, până când rezultatele cercetărilor ulterioare îi vor furniza un răspuns mai tranşant la întrebarea sa.

Tabelul 10.1 Valori Critice pentru testul *t*

	Nivele de semnificaţie pentru testul direcţionat (one-tailed)					
	0.10	0.05	0.025	0.01	0.005	0.0005
	Nivele de semnificaţie pentru testul nedirecţionat (two-tailed)					
df	0.20	0.10	0.05	0.02	0.01	0.001
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.689
3	1.638	2.353	3.182	4.541	5.841	12.941
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.859
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.405
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.767
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.045	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646
40	1.303	1.684	2.021	2.423	2.704	3.551
60	1.296	1.671	2.000	2.390	2.660	3.460
∞	1.282	1.645	1.960	2.326	2.576	3.291

Valoarea t obţinută din calcule este semnificativă dacă ea este mai mare sau egală cu valoarea listată în tabel

Testul *t* pentru eşantioane mari

In cazul in care cele doua eşantioane sunt mari, adică depăşesc limita convenţională a celor 30 indivizi, eroarea standard a diferentelor dintre medii se calculează după o formula mai simplă în care nu intervin gradele de libertate, iar testul care se utilizează este testul *z* (Vezi capitolul despre distribuţia *z* din manualul Statistica sociala si analiza datelor I). Indicatorul *z* are o distribuţie normală cu media 0 si abaterea standard 1 si nu depinde de numărul gradelor de libertate (important e ca eşantionul sa fie mare). Formula de calcul a statisticii t (in acest caz echivalent cu indicatorul z) este:

$$t = \frac{m_1 - m_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Testul *t* pentru eşantioane perechi

Testul *t* se poate utiliza pentru evaluarea de pildă a impactului unui program de intervenţie asupra unei trăsături a populaţiei cuprinse în eşantion (atitudine, comportament, caracteristică psiho-socială, etc.). Valoarea lui *t* se calculează pe baza mediei diferenţelor dintre valorile variabilei respective înainte de intervenţie, respectiv după intervenţie. Valoarea lui *t* exprimă „succesul" intervenţiei, adică măsura în care, în medie, valorile variabilei respective s-au modificat. Modul de calcul diferă însă faţă de cel de la eşantioanele independente, dar programul SPSS tratează, din fericire, toate variantele de calcul.

$$t = \frac{m_D}{s_D / \sqrt{n}}$$

unde m_D este media distribuţiei D (a diferenţelor dintre cele două măsurări), iar s_D este abaterea standard a distribuţiei D (măsoară împrăştierea distribuţiei D).

Exemplu: Doriţi să aflaţi dacă în urma participării la un grup de suport pentru mamele singure, nivelul de stres al celor 8 participante s-a redus. In acest caz ipoteza de cercetare este ca „participarea la grupul de suport reduce semnificativ nivelul de stres al mamelor singure". Aplicând acelaşi chestionar care măsoară nivelul de stres înainte şi după 6 luni de participare la grupul de suport, obţineţi rezultatele din tabelul următor. Aplicaţi testul t, situându-vă la un nivel de semnificaţie de .0.01.

	Înainte de program (X_1)	**După program (X_2)**	**D ($X_2 - X_1$)**	**D - m_D**	**$(D - m_D)^2$**
	7	7	.00	.50	.25
	9	8	-1.00	-.50	.25
	11	12	1.00	1.50	2.25
	10	9	-1.00	-.50	.25
	6	6	.00	.50	.25
	7	6	-1.00	-.50	.25
	12	11	-1.00	-.50	.25
	6	5	-1.00	-.50	.25
$\sum X$	68	64	-4		$\sum (D - m_D)^2 = 4$
n	8	8	8		
$m = \frac{\sum X}{n}$	8,50	8	m_D= - 0,5		

$s_D = \sqrt{\frac{(D - m_D)^2}{n-1}}$					$\sqrt{\frac{4}{7}} = 0,75$

Introducem valorile astfel obținute în formula de mai sus şi obținem:

$$t = \frac{-0,5}{0,75 / \sqrt{8}} = -2,08$$

De reținut: În privința procedurii statistice, nu contează care diferență o utilizăm, $X_1 - X_2$ sau $X_2 - X_1$. Depinde de ceea ce dorim să determinăm. Important este ca, în final, să interpretăm corect rezultatul obținut, ținând cont de semnul diferenței şi de semnificația concretă a acestuia. Totuşi, în modele de tipul „înainte / după" (tratament, intervenție), este mai potrivită utilizarea diferenței $X_2 - X_1$.

Interpretare: Comparăm t calculat (în valoare absolută) cu valoarea critică la nivelul de semnificație 0.01, pentru o ipoteză nedirecționată (nu am specificat direcția legăturii dintre variabile) şi la 7 grade de libertate (*df* = n - 1 = 8 - 1 = 7) şi observăm că 2.08 < 3.49. Deoarece probabilitatea de a obține un nivel de stres mai redus ca urmare a şansei este mai mare decât nivelul pe care ni l-am propus drept criteriu de decizie (adică mai mică de 1%), acceptăm ipoteza nulă. În acest caz, datele noastre nu sprijină ipoteza cercetării, ca urmare nu putem accepta că efectul obținut se datorează participării la grupul de suport. În concluzie, grupul de suport nu reduce în mod semnificativ nivelul de stres al participantelor.

10.4. Prezentarea statisticii *t*

De obicei este util a prezenta mărimile brute ale cazurilor în ambele grupuri (dacă eşantionul nu este prea mare) astfel ca cititorii să poată face propriile comparări. Dacă nu dorim să raportăm rezultatele individuale, vom prezenta cel puțin rezultatele medii pentru fiecare grup. Trebuie reținut că diferența actuală între medii ajută cititorul în a determina dacă o legătură semnificativ statistică este în mod real un rezultat esențial. Astfel vor mai fi prezentate pe lângă valoarea lui *t,* gradul de libertate (*df*) şi valoarea probabilității *p* asociate. Dacă se realizează mai multe teste *t*, este folositor un singur tabel care să conțină toate rezultatele obținute. Dacă sunt calculate doar puține teste *t*, rezultatele sunt de obicei prezentate în corpul textului, mai degrabă decât într-un tabel.

În majoritatea studiilor statistice, se raportează atât analizele care nu au permis respingerea ipotezei nule, cât şi cele care au permis-o. În unele cazuri este suficient să se raporteze că nu s-a găsit că variabilele ar fi legate la un nivel statistic semnificativ. În situațiile în care ipoteza nulă este ipoteza centrală a cercetării (adică, încercăm să dovedim ca falsă o legatură care în mod obişnuit este considerată ca fiind adevărată), suportul statistic sau nesprijinirea pentru respingerea ipotezei nule trebuie, firește, să fie prezentat.

10.5. Când nu se foloseşte testul *t*

Ca şi testul hi-pătrat, popularitatea testului *t* printre asistenții sociali ne poate conduce la utilizări greşite. În graba noastră de a folosi o statistică care ne este familiară, putem folosi câteodată testele *t* în situații în care ele nu sunt corespunzătoare şi în care ar fi putut fi utilizate alte teste mai potrivite. De obicei când testul *t* este folosit greşit avem una din următoarele situații: (1) când ignorăm forma distribuției pentru variabila de nivel interval în populație şi (2) când nu suntem atenți la analizarea corectă a informațiilor, încercând să prezentăm un rezultat şocant.

Forma distribuției

Faptul că variabila independentă este de nivel interval şi că cea dependentă este de nivel nominal nu justifică folosirea lui *t* pentru a determina prezența unei legături între două variabile. Testul *t* este un test din categoria celor parametrice, ceea ce înseamnă că este proiectat pentru a fi folosit doar când variabila de nivel interval poate fi considerată a fi normal distribuită în populație. Dacă distribuția de frecvențe pentru valorile variabilei în populație este vizibil asimetrică, va trebui folosit un alt test, ca de pildă, testul Mc

Neamar, testul exact Ficher, testul Man-Whitney, etc. Credibilitatea rezultatelor cercetării este serios periclitată dacă testul *t* este folosit cu variabile de nivel interval sau rapoarte care nu se prezintă ca o curbă normală.

Abordarea şocantă

O a doua greşeală obişnuită implică calcularea în serie a testelor *t*, cu o singură variabilă dependentă legată eventual de mai multe variable independente. În câteva exemple particulare ale acestei erori, cercetătorii au calculat testele *t* folosind sute de *posibile* variabile independente (fără prea mare legătură), doar pentru a anunța cu mândrie că ei au găsit o legătură semnificativ statistică între una dintre variabilele independente şi variabila dependentă.

Teoria probabilităților sugerează că astfel de rezultate reflectă probabil erori de Tipul I şi că legătura aparentă va fi legată de şansă şi nimic mai mult. Conform legilor probabilității, o variabilă dependentă va părea că are o asociere semnificativ statistică cu câteva variabile independente dacă sunt testate suficiente legături. În situațiile *unde* există *motive* să credem că mai multe variabile independente diferite pot fi în legătură cu variabila dependentă, ar trebui să folosim alte teste statistice, special proiectate pentru astfel de situații.

10.6. Un exemplu

Descrierea situației care generează un studiu

Florin este un asistent social într-un centru de planificare familială. El a fost învățat în agenția sa că cea mai bună formă pentru consilierea maritală este de a întâlni ambii parteneri (soțul şi soția) împreună. Înainte cu cinci ani, a tratat 20 de cupluri în care consilierea nu s-a putut realiza decât individual (soț sau soție), din cauza programului pe care îl aveau aceştia la serviciu. El a fost surprins observând că, deşi ei nu au consiliați niciodată împreună, după primele interviuri, toate cele 20 de cupluri păreau să realizeze un progres excelent în rezolvarea problemelor lor maritale.

De-a lungul a câtorva ani, Florin a consiliat tot mai multe cupluri pe o bază individuală. Crezând că a obținut rezultatele bune datorită metodei folosite, el a încurajat şase dintre colegii săi să consilieze cuplurile cu dificultați maritale, întâlnindu-i separat mai degrabă decât împreună. Alți asistenți sociali au fost de asemenea surprinşi de progresele excelente ale clienților lor. Florin nu a fost pregătit să concluzioneze că metoda de consiliere maritală individuală este preferabilă consilierii cuplului. El a decis să realizeze o mică cercetare pentru a vedea dacă poate găsi suport statistic pentru ipoteza că satisfacția maritală este cel mai bine sporită când partenerii sunt tratați în consiliere individuală şi nu de cuplu.

Ipotezele care vor fi testate

A început să caute în literatura de specialitate de asistență socială şi a găsit suport pentru afirmația conform căreia consilierea în cuplu este mai adecvată în obținerea satisfacției maritale, decât cea individuală. Dar cum s-a aventurat în literatura de specialitate din alte domenii ca psihologia şi consilierea pastorală, Florin a găsit o cantitate mare de suport pentru ideea că satisfacția maritală ar fi mult mai probabil să rezulte din consilierea individuală. Explicația cea mai uzitată era că persoanele consiliate tind să discute despre insatisfacție mai uşor şi mai cinstit atunci când celălalt soț nu este prezent. Florin a concluzionat că literatura de specialitate era în contradicție. Totuşi, el a simțit că propriile observații şi cele ale colegilor săi au fost suficiente pentru a justifica o ipoteză direcțională:

Printre clienții care a beneficiat de consiliere maritală, aceia care au fost consiliați individual vor reflecta un nivel mai înalt de satisfacție maritală decât cei care au fost consiliați în cuplu.

O privire asupra metodologiei

Florin a realizat o mică cercetare pentru a-şi testa ipoteza. El a primit permisiunea din partea directorului agenției de a-i fi repartizați la întâmplare clienții noi care au cerut consiliere maritală pe o perioadă de trei luni pentru a realiza consiliere individuală sau de cuplu. Pentru a realiza cercetarea, clienții au fost repartizați la întâmplare în două grupuri de subiecți. Toți cei şase asistenți sociali care au folosit

anterior consilierea individuală (şi care, deci, aveau experienţă) au participat ca şi consilieri la studiul lui Florin. Începând cu luna următoare, fiecare cuplu care era de acord a fost repartizat la unul dintre cei şase asistenţi sociali pentru a participa împreună la o consiliere de 50 de minute pe săptămână; cuplurile care au rămas au fost repartizate pentru consiliere individuală, timp de 25 de minute în fiecare lună. Cei care nu au fost de acord cu acest aranjament au fost de asemenea trataţi, dar nu au mai fost incluşi ca subiecţi în acest studiu.

Metoda de consiliere (individuală sau în cuplu) a fost variabila independentă. S-a decis că variabila dependentă, satisfacţia maritală, va trebui măsurată dupa zece săptămâni consecutive de consultanţă. A fost aleasă o scală pentru a măsura gradul satisfacţiei maritale, ea fiind adaptată pentru a produce informaţii de nivel interval.

Paisprezece cupluri au fost văzute în sesiuni de consiliere individuală şi paisprezece au fost văzuţi ca şi cuplu. Toţi clientii au completat scala satisfacţiei maritale. Florin a comparat mărimea mediei scalei pentru clienţii care au fost văzuţi individual (grupul experimental) cu mărimea mediei scalei pentru clienţii care au fost văzuţi ca şi cuplu (grupul de control). Variabila "satisfacţie maritală" măsurată pe scală a fost găsită ca fiind normal distribuită, deci Florin s-a simţit în măsură să folosească testul *t* pentru analiza statistică. El a încercat să determine dacă diferenţa dintre mărimile mediei (pentru cele două grupuri) a fost suficient de mare pentru a-i permite să respingă ipoteza nulă. El a dorit să fie capabil să concluzioneze că o legătură reală dintre două variabile a fost probabil explicaţia pentru diferenţele observate în eşantion.

Rezultatele

Valoarea lui *t* pentru informaţiile lui Florin, a fost de 1.312 . Din tabelul valorilor critice ale lui *t* (Tabelul 10.1), el a observat că ar fi avut nevoie de o valoare minimă a lui *t* de 1.706 pentru a putea respinge ipoteza nulă (pornind de la o ipoteză direcţională şi folosind o statistică semnificativă la nivelul 0.05, un eşantion de 28 indivizi statistici şi 26 de grade de libertate).

Interpretarea rezultatelor şi tragerea concluziilor

Florin a notat că dacă ar fi respins ipoteza nulă pe baza informaţiilor lui, ar fi desconsiderat mai mult decât 1 din 10 (10 procente) şanse în comiterea unei erori de Tip I. El avea în mod clar o lipsă de suport statistic pentru ipoteza sa. Dezamăgirea lui iniţială a fost chiar şi mai mare atunci când a privit mărimile mediilor celor două grupuri. Clienţii care au participat la consiliere individuală au avut media cu ceva mai mică decât cei care au fost consiliaţi împreună. Florin a privit rezultatele sale cu mai multă obiectivitate. Astfel, el a realizat că lipsa argumentului statistic pentru legătura dintre metoda de consiliere şi satisfacţia maritală poate fi un rezultat folositor pentru sine însuşi. Imposibilitatea de a respinge ipoteza nulă poate fi interpretată în termenii existenţei unei diferenţe prea mici determinate de metoda de consiliere folosită.

Florin şi-a pus, de asemenea, problema cum de a putut greşi. Rezultatele cantitative din studiul său au fost inconsistente în ceea ce priveşte impresiile anterioare. El s-a întrebat dacă el sau alţi asistenţi sociali nu cumva au *perceput* doar că consilierea individuală făcută de ei duce la rezultate superioare, pentru că ei au fost surprinşi că ceilalţi clienţi consiliaţi în cuplu au putut obţine rezultate superioare. Firește el s-a întrebat dacă ipoteza sa poate fi încă corectă. S-ar putea ca adevărata legătură dintre variabila dependentă şi cea independentă să fi fost ascunsă de tehnica măsurătorilor sau de influenţa altor variabile (de exemplu, experienţa mare a asistenţilor sociali care au efectuat consilierea în cuplu). Florin a concluzionat că sunt necesare studii suplimentare pentru a clarifica această situaţie.

10.7. Folosirea SPSS in compararea mediilor

Vom face o aplicaţie SPSS pentru acest test, folosind fişierul de date „1991 US General Social *Surv*ey". Dorim să testăm dacă există o diferenţă semnificativă între femei şi bărbaţi în ceea ce priveşte nivelul educaţional atins, măsurat în numărul total de ani de studiu absolviţi. Nivelul educaţional al respondenţilor este dat de variabila *educ*. Alegem secvenţa de opţiuni: ***Statistics – Compare Means – Independent Sample T Test***. Ultima opţiune deschide o fereastră de dialog în care specificăm variabila ale cărei medii dorim să le testăm (*Test Variable – educ*) şi variabila care ne dă cele două subpopulaţii – în acest caz sexul (*Grouping Variable, Define groups*, în care trecem codurile celor două sexe: 1 pentru

bărbaţi şi 2 pentru femei). De asemenea, la *Options* specificăm nivelul de încredere dorit (vom folosi 95%). Iată ce ne afişează programul SPSS:

```
T-TEST
  GROUPS=sex(1 2)
  /MISSING=ANALYSIS
  /VARIABLES=educ
  /CRITERIA=CIN(.95) .
```

T-Test

Group Statistics

	Respondent's Sex	N	Mean	Std. Deviation	Std. Error Mean
Highest Year of School Completed	Male	633	13.23	3.143	.125
	Female	877	12.63	2.839	.096

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Differ-ence	Std. Error Differ-ence	95% Confidence Interval of the Difference	
									Lower	Upper
Highest Year of School Completed	Equal variances assumed	11.226	.001	3.887	1508	.000	.60	.155	.298	.906
	Equal variances not assumed			3.824	1276.5	.000	.60	.157	.293	.911

Ca de obicei, apare sintaxa corespunzătoare opţiunilor din meniu folosite. În tabelul următor apar valorile mediilor variabilei „nivelul educaţional al respondentului" pentru cele două populaţii (din eşantion), abaterile standard şi erorile standard corespunzătoare. Ultimul tabel este cel care prezintă rezultatele testului de semnificaţie. În primele două coloane avem valori pentru testarea egalităţii varianţelor distribuţiilor variabilei pentru cele două populaţii, sau altfel spus, similaritatea omogenităţii celor două distribuţii, folosind testul F[3]. Aici cele două distribuţii sunt semnificativ diferite (p=0.001 sau P=99.9%); în consecinţă, vom citi valorile de pe rândul de jos, calculate pentru varianţe semnificativ diferite ale celor două distribuţii. Valoarea lui t este 3.824, la nivelul de semnificaţie p=0.000, ceea ce înseamnă că sunt 99.9% şanse ca diferenţa să fie reală. De asemenea, dacă ne uităm la intervalul de confidenţă pentru diferenţa mediilor – *Confidence Interval of the* (Mean) *Difference*, observăm că nu conţine valoarea 0, deci nu este posibil ca diferenţa mediilor să fie 0, adică mediile să fie egale. Aşadar, există o diferenţă semnificativă între femei şi bărbaţi, la nivelul populaţiei, în ceea ce priveşte numărul de ani de şcoală absolviţi (bărbaţii au, în medie, mai mulţi ani de şcoală decât femeile: mediile lor sunt 13.23, respectiv 12.63).

[3] Puteţi afla mai multe despre testul F din lucrarea „Metode statistice aplicate în ştiinţele sociale" (2000), Rotariu T. (coord.), Ed. Polirom, Iaşi, pp. 102-106.

Observţie: În cazul eşantioanelor perechi, procedura SPSS este următoarea: din meniul ***Statistics*****,** ***Compare Means*****,** ***Paired-Sample T test***. Introducem variabilele pe care le testăm în rubrica *Grouping Variables*.

Întrebări pentru studiu

1. Care este combinaţia de nivele de măsurare a variabilelor utilizate în testul *t*?
2. De ce dimensiunile cerute de testul *t* pentru volumul eşantionului şi a subeşantioanelor îl fac ideal pentru cercetările în asistenţă socială?
3. Dacă ipoteza nulă este corectă, valoarea medie a variabilei din eşantion va fi similară sau foarte diferită de valoarea medie a variabilei din celălalt eşantion?
4. Cum se determină gradele de libertate pentru statistica *t,* într-un set de date concrete?
5. Ce pas suplimentar este necesar în a determina dacă o valoare *t* semnificativă din punct de vedere statistic oferă suport pentru o ipoteză direcţionată?
6. Cum poate un test *t* care nu a găsit o valoare semnificativă statistic să reflecte un rezultat care să fie util pentru un practician?

Exerciţiu pentru acasă 9

Pornind de la fişierul de date „1991 US General Social Survey", testaţi folosind programul SPSS, dacă există o diferenţă semnificativă între femei şi bărbaţi în privinţa prestigiului ocupaţional (variabile sex si prestg80). Comentaţi rezultatele obţinute.

Modul 11

Corelația

Acest capitol explică modul în care putem să analizăm legătura dintre două variabile de nivel interval sau rapoarte prin procedura *analizei de corelație*.

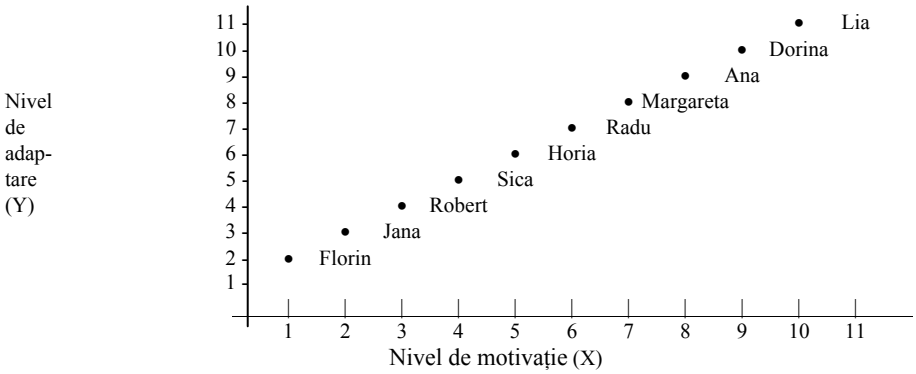
11.1. Conceptul de corelație

Un exemplu de legătură între două variabile poate fi cel al nivelului de adaptare socială a persoanelor considerate ne-integrate social după nivelul de motivație pentru tratament. Tabelul 11.1 cuprinde datele culese de la zece clienți cu tulburări de comportament. Pentru fiecare client avem atât valoarea nivelului de motivație (variabila cantitativă X) cât şi valoarea nivelului de adaptare socială (variabila Y, tot de nivel interval). Ambele variabile se presupune că au fost măsurate pe o scală de la 0 la 11 puncte. O legătură între aceste două variabile este evidentă deoarece, fără excepție, nivelul înalt al motivației pentru tratament este asociat cu nivelele înalte de adaptare socială şi viceversa. Florin de exemplu, a înregistrat cea mai scăzută valoare la ambele teste: 1 punct la nivelul de motivație şi 2 puncte la nivelul de adaptare socială; Jana se situează pe penultimul loc, cu cele 2, respectiv 3 puncte obținute pe scala nivelurilor de motivație si de adaptare socială., în timp ce Lia a atins nivelurile cele mai înalte pentru ambele variabile (rezultatele 10 şi respectiv 11).

Tabelul 11.1 Rezultate înregistrate de 10 clienți cu probleme de adaptare

Numele clientului	Nivel de motivație (X)	Nivel de adaptare socială (Y)
Florin	1	2
Jana	2	3
Robert	3	4
Sica	4	5
Horia	5	6
Radu	6	7
Margareta	7	8
Ana	8	9
Dorina	9	10
Lia	10	11

Figura 11.1 Legătură perfectă pozitivă între două variabile: nivel de adaptare socială şi nivel de motivație (din Tabelul 11.1)



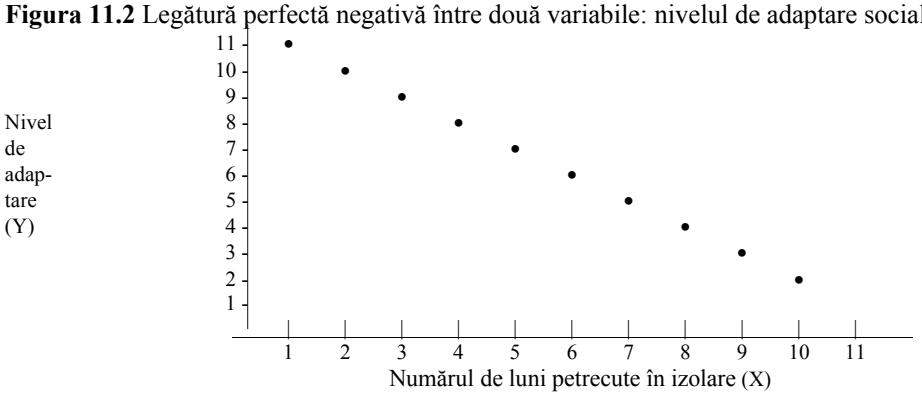
Această legătură poate fi vizualizată cu ajutorul diagramei de împrăştiere, ca cea ilustrată în Figura 11.1. Axa orizontală reprezintă rezultatele individuale obținute de clienți la testul care măsoară nivelul de motivație (X), în timp ce axa verticală reprezintă rezultatele individuale privind nivelul de adaptare socială

(Y). Fiecare punct reprezintă un caz şi totodată o pereche de valori, măsuri ale celor două variabile pentru client. Dacă punctele se conectează, ele vor forma o linie dreaptă, indicând că cele două variabile sunt perfect corelate. Asemenea perfecţiune este rar întâlnită în practica cercetării de asistenţă socială. Aici ea este folosită pentru a ilustra conceptul de corelaţie.

Intensitate şi direcţie

Figura 11.1 pune în evidenţă două aspecte importante ale relaţiei dintre variabile: intensitatea şi direcţia. In ceea ce priveşte ***intensitatea***, o legătură între două variabile este cea mai puternică atunci când rezultatele tuturor clienţilor cad pe o linie drepte. Linia care trece prin toate punctele (reprezentând clienţii) diagramei de împrăştiere este numită *linie de regresie*. În aceste, foarte rare, cazuri în care există o *legătură perfectă*, putem prezice cu precizie maximă (de 100%) că unei valori Y îi corespunde o anume valoare X, şi viceversa. În cazurile obişnuite unde intensitatea legăturii este mai puţin perfectă, linia de regresie este mai puţin distinctă, iar capacitatea noastră de a prezice valorile unei variabile din valorile alteia este supusă erorii.

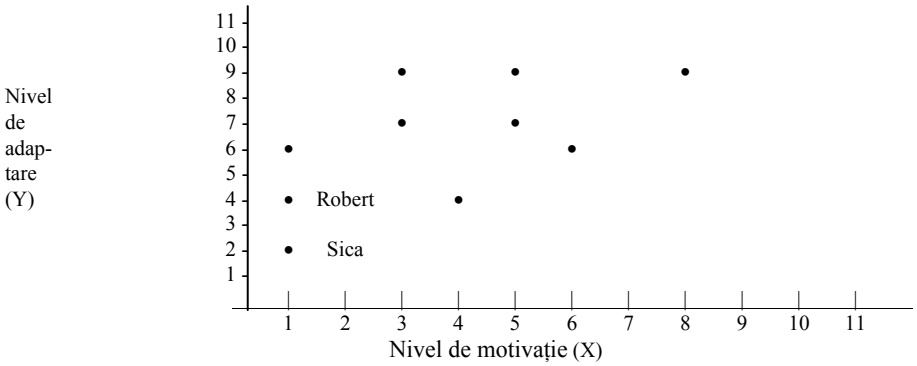
Figura 11.2 Legătură perfectă negativă între două variabile: nivelul de adaptare socială şi durata izolării



În afară de intensitate, corelaţia are şi un al doilea aspect: ***direcţia*** legăturii. Legătura din Figura 11.1 este una pozitivă: legătura dintre nivelul de motivaţie şi nivelul de adaptare socială a clienţilor este o *legătură pozitivă* (o legătură perfectă pozitivă). Unor valori înalte pentru X ii corespund valori înalte pentru Y, şi viceversa. Într-o *legătură negativă* (ca cea din Figura 11.2), valorile înalte ale unei variabile corespund unor valori scăzute pentru cea de-a doua variabilă. Corelaţia perfectă pozitivă şi corelaţia perfectă negativă sunt foarte rare în realitatea socială, ele apărând mai des în domeniul fizicii.

În cercetările sociale, putem întâlni multe variabile care sunt corelate (pozitiv sau negativ) dar există situaţii în care nu există legaturi de corelaţie între două variabile. Acestea lucru este susţinut de ipoteza nulă. Totuşi majoritatea legăturilor dintre variabilele de nivel interval sau rapoarte reflectă un anumit grad de corelaţie, mergând de la legături aproape perfecte şi ajungând până la legături abia distinctibile numite şi *corelaţii nonperfecte*.

Figura 11.3 Legătură nonperfectă pozitivă între două variabile: nivelul de motivaţie şi nivelul de adaptare socială



Similar, figura 11.4 furnizează un exemplu de legătură negativă care este mai puțin perfectă. Figura 11.5 prezintă un exemplu în care nu se poate observa nici un fel de legătură între cele două variabile.

Figura 11.4 Legătură nonperfectă negativă între două variabile: nivelul de motivație şi nivelul de adaptare socială

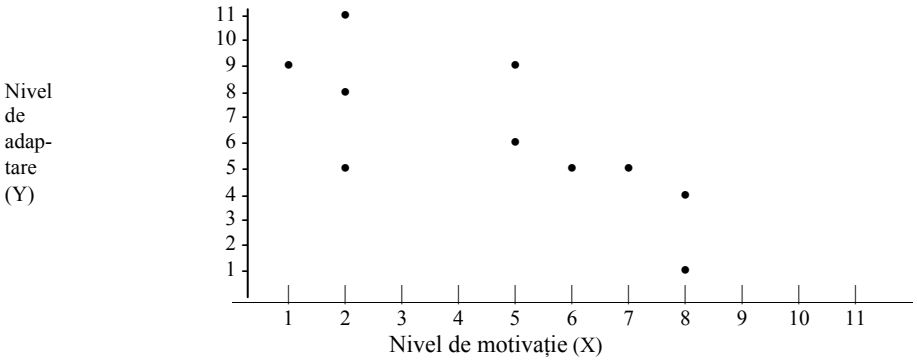
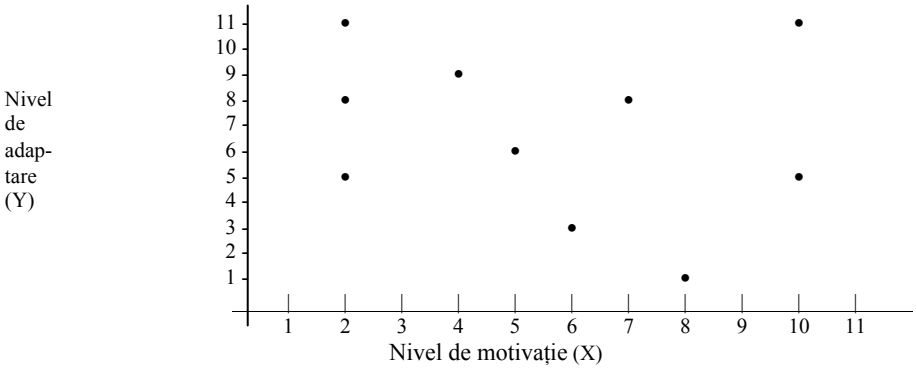


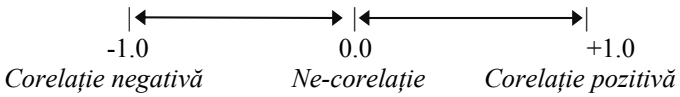
Figura 11.5 Nici o legătură între două variabile: nivelul de motivație şi nivelul de adaptare socială



11.2. Coeficientul de corelație şi proprietățile lui fundamentale

Figurile 11.1 - 11.5 sunt diagrame de împrăştiere standard. Ele sunt modalități dificile şi consumatoare de timp de prezentare a perechilor de valori pentru un număr mare de cazuri. O cale mult mai eficientă de punere în evidență a unei relații dintre variabile de nivel interval sau rapoarte este folosirea unei indicator statistic numit ***coeficient de corelație***. Acesta furnizează o valoare numerică care indică atât intensitatea cât şi direcției unei legături. Aşa cum este prezentat în Figura 11.6, coeficientul de corelație variază într-un continuu de valori, de la o extremă -1,0 (legătura perfect negativă) la o altă extremă 1.0 (legătura perfect pozitivă), cu 0.00 (ne-corelație) ca punct de mijloc. Un coeficient de corelație nu poate fi mai mare decât 1.00 sau mai mic decât -1.0.

Figura 11.6 Intervalul în care variază coeficientul de corelație



Cu cât valoarea numerică a coeficientului de corelație este mai aproape de una din valorile extreme (-1 sau -1), cu atât mai puternică este legătura dintre cele două variabile. De exemplu, un coeficient de 0,92 este mai apropiat de o corelație perfectă decât oricare din coeficienții -0,65 sau 0,60 şi de aceea, sugerează o mai puternică corelație decât oricare dintre cele două. Cu cât coeficientul este mai apropiat de mijlocul acestui interval cu atât legătura dintre cele două variabile este mai slabă. Un coeficient de corelație care este apropiat de valoarea 0 sugerează că nu există nici o legătură de corelație între cele două variabile.

Semnul plus sau minus indică direcția legăturii. De exemplu, corelația dintre nivelul îndemânării asistenților sociali şi anii de experiența profesională în asistență socială poate fi de 0,85. Absența semnului minus indică o legătură pozitivă. Corelația dintre nivelul de îndemânare şi nivelul de apatie a celor care muncesc poate fi de -0,75, indicând astfel o puternică legătură negativă. Există persoane care sunt mai apatice şi mai ne-implicate în munca lor, astfel este normal ca ele să fie mai puțin îndemânatice, şi viceversa.

Pot fi utilizate o varietate de analize de corelație. Una din procedurile cele mai frecvent utilizate este şi cea care va fi prezentată în acest capitol. Este vorba de ***coeficientul de corelație Pearson*** (sau coeficientul de corelație ***r***), (sau *corelația produsului momentelor*) pe care îl vom numi în discuțiile ce urmează, simplu: *r*. El se calculează cu formula:

$$r = \frac{\sum (x_i - m_x) * (y_i - m_y)}{N * \sigma_x * \sigma_y}$$

unde:

x_i şi y_i reprezintă valorile individuale ale distribuțiilor X şi Y
m_x şi m_y reprezintă mediile distribuțiilor X şi Y
σ_x şi σ_y reprezintă abaterile standard ale distribuțiilor X şi Y
N este volumul eşantionului

Suma produselor perechilor $(x_i-m_x)*(y_i-m_y)$ împărțită la volumul eşantionului este cunoscută sub denumirea de *covarianță* (notată cu ***cov(X,Y)***) sau *produsul momentelor de ordinul întâi*. Coeficientul *r* se obține prin împărțirea covarianței la produsul dintre abaterile standard ale celor două variabile. Procedând astfel, se scapă de unitățile de măsură ale covarianței şi de intervalul ei nedefinit de variație, obținându-se un indice (*r*) care are proprietăți mult mai bine definite.

Înțelegerea indicatorului statistic *r*

Folosind *r*, încercăm să determinăm statistic ceea ce arată o diagramă de împrăştiere - adică, până la ce grad punctele reprezentând valorile celor două variabile pentru fiecare caz vor tinde să formeze o linie dreaptă. Dacă valoarea lui *r* este mare (dacă se apropie de 1,00 sau -1,00) punctele vor tinde să se "strângă" aproape de linia de regresie; iar dacă valoarea lui *r* este mică (în apropierea lui 0,00) o diagramă a punctelor va reflecta distanțarea punctelor față de dreapta de regresie.

Pentru a înțelege cum a apărut formula lui *r* ca un indicator al corelației dintre variabile, este bine să ținem minte că *r* este interpretat ca o expresie a abilității noastre de a prezice o valoare a unei variabile din informațiile pe care le avem despre o valoare a celeilalte variabile. Nu este implicată nici o relație de cauzalitate. Argumentul care explică ce variabilă cauzează variația alteia, trebuie să se bazeze pe alți factori.

Trebuie remarcat în formula lui *r,* că numărătorul este derivat din perechile de valori (una pentru fiecare variabilă) ale fiecărui caz. Suma de la numărător reflectă gradul în care variază împreună cele două variabile iar prin împărțirea la volumul eşantionului avem expresia variației medii. Exprimat în alt fel, covarianța ne spune în ce grad anumite valori ale unei variabile tind să fie găsite împreună cu anumite valori ale celeilalte variabile.

La numitor, formula introduce cantitatea totală de variație a tuturor datelor; anume produsul abaterilor standard al celor două variabile. Valoarea *r* care rezultă din formula finală, este proporția din variația totală ce revine covariației de la numărător. Într-o corelație perfectă (r = 1,0 sau r = -1,0) 100 % din variația totală a celor două variabile va fi explicată de către covariația de la numărător. Invers, în acele cazuri rare în care nu există corelație (r = 0,0) între variabile, nimic din variația totală nu va fi explicată de covariația rezumată în numărător. Cunoscând valoarea pentru o variabilă a unui caz în care r = 0,0 ea nu ne va ajuta cu nimic în a prezice valoarea pentru cea doua variabilă pentru acel caz.

Valoarea indicatorului statistic *r* care rezultă din formulă ne spune mai multe. Dacă ridicăm la pătrat valoarea *r*, noua valoare r^2 va fi proporţia variaţiei dintr-o variabilă care poate fi "explicată" prin legătura ei cu cealaltă variabilă. Ţineţi minte că termenul *explicaţie* pe care îl folosim aici, nu înseamnă *cauzalitate*. Se referă doar la abilitatea noastră de a prezice valorile unei variabile dacă ştim valorile altei variabile. De exemplu, un coeficient *r* de 0,80, relativă la corelaţia dintre numărul de interviuri de tratament şi la rezultatele obţinute la testul stimei de sine, înseamnă că 64% (r^2=0,80 x 0,80 = 0,64) din variaţia rezultatului obţinut la testul stimei de sine poate fi explicat pe baza numărului de interviuri de tratament. Celelalte 36 de procente ale variaţiei valorilor stimei de sine (1,00 - 0,64) rămân ne-explicate, sau factorii care explică acestea sunt altele. Capacitatea noastră de a prezice "imaginea de sine" a clienţilor dacă ştim numărul de interviuri avute ar trebui să fie bună, dar cu siguranţă nu poate fi perfectă. Doar un *r* de +1,0 sau de -1,0 (o corelaţie perfectă) ne-ar permite să facem preziceri cu o acurateţe de 100% ($1,0^2 = 1,0$).

Coeficientul de corelaţie *r* este un indicator al intensităţii şi al direcţiei unei legături care există între valorile unei variabile şi valorile altei variabile. Aşa cum am subliniat în prezentarea lui hi-pătrat, folosim greşit un test când încercăm să mergem mai departe decât este principala sa destinaţie (predicţia în cazul lui *r*) şi încercăm să îi atribuim mai multe atribuţii (implicaţii în cauzalitate) decât este el capabil să realizeze. Exemplul şi aplicarea pas cu pas a formulei care urmează, va sublinia faptul că *r* (la fel ca alte teste statistice) este doar o manevrare matematică a numerelor care pot furniza doar un răspuns parţial pentru întrebările cercetării noastre. Răspunsurile plenare trebuie să vina din logică, teorie, intuiţie, experienţă practică, sau din alte domenii, mai puţin din procesele "mecanice".

Calcularea indicatorului statistic *r*

Pentru calculul coeficientului de corelaţie *r* să pornim de la un exemplu.

Exemplu: Coordonatorul unui centru de servicii familiale doreşte să realizeze un plan de perfecţionare a asistenţilor sociali angajaţi în scopul îmbunătăţirii calităţii serviciilor oferite clienţilor. Înainte de a organiza activitatea de perfecţionare, va dori să afle dacă există o legătură între numărul de ore de specializare urmate de asistenţii sociali pe durata ultimilor 5 ani şi numărul de eşecuri în intervenţie raportat de asistenţii sociali cu ocazia autoevaluărilor anuale în aceeaşi perioadă. Ipoteza de la care porneşte coordonatorul este că cu cât creşte numărul de orelor de perfecţionare ale asistenţilor sociali cu atât scade numărul de eşecuri în intervenţie. În urma investigaţiei, el obţine următoarele date:

Tabel 11.2. Date despre asistenţii sociali din eşantion

Identificator asistent social	A.s. 1	A.s. 2	A.s. 3	A.s. 4	A.s. 5	As. 6	A.s. 7
Nr. ore de perfecţionare	20	30	30	40	50	50	60
Nr. de eşecuri în intervenţie	7	6	5	4	3	2	1

Coeficientul de corelaţie măsoară intensitatea unei relaţii liniare dintre două variabile X şi Y, dintre care una poate fi considerată variabilă independentă (în acest caz, numărul de ore de perfecţionare), iar cealaltă variabila dependentă (în acest caz, numărul de eşecuri în intervenţie)[4]. Semnul coeficientului de corelaţie *r* indică direcţia legăturii dintre cele două variabile.

Pornind de la datele empirice se poate construi Tabelul 11.3.

Concluzii: Coeficientul de corelaţie între numărul de ore de formare realizate şi numărul de eşecuri în intervenţie este -.98, foarte apropiat de -1, indicând o corelaţie negativă foarte puternică. Pe baza acestui rezultat, Robert concluzionează că asistenţii sociali care s-au perfecţionat mai intens au mai puţine eşecuri în munca de intervenţie. Totuşi, este important de reţinut: corelaţia nu implică şi cauzalitate. Aşadar, nu se poate afirma faptul că un anumit număr de ore de formare cauzează un anumit număr de eşecuri în intervenţie. Există şi alţi factori cu care cele două variabile sunt, la rândul lor, corelate (cum ar fi de pildă, trăsăturile de personalitate ale asistenţilor sociali). Dar putem oare extinde rezultatul la întreaga populaţie a asistenţilor sociali?

[4] Nu este absolută nevoie să facem distincţie între variabila independentă şi cea dependentă, pentru că *r* nu cere acest lucru.

Tabelul 11.3. Schemă de calcul pentru coeficientul de corelație *r* a lui Pearson. Corelatia dintre numarul de ore de perfectionare (X) si numarul de cazuri esuate (Y), la 7 asistenti sociali

x	y	$x-m_x$	$(x-m_x)^2$	$y-m_y$	$(y-m_y)^2$	$(x-m_x)(y-m_y)$
20	7	-20	400	3	9	-60
30	6	-10	100	2	4	-20
30	5	-10	100	1	1	-10
40	4	0	0	0	0	0
50	3	10	100	-1	1	-10
50	2	10	100	-2	4	-20
60	1	20	400	-3	9	-60
∑= 280	**∑= 28**	**∑= 0**	**∑= 1200**	**∑= 0**	**∑= 28**	**∑= -180**
m_x = 40	**m_y = 4**		**σ^2_x= 171,4**		**σ^2_y = 4**	**cov(x,y)= -25,7**
			σ_x = 13,1		**σ_y = 2**	

$r = \text{cov}(X, Y) / (\sigma_x * \sigma_y) = -25,7 / (13,1 * 2) = -0,98$
test direcțional, $n = 7$, $p < 0.005$

Testarea semnificației statistice a lui *r*

Cât de mare trebuie să fie *r* pentru a putea respinge în siguranță ipoteza nulă şi să avem suport statistic pentru existența unei legături reale între cele două variabile? Ca şi la alte legături bivariate, coeficientul de corelație este gândit să fie suficient de puternic ca să ajungă la sau să depăşească nivelul predeterminat al semnificației statistice. Nivelul de semnificație de 0,05 justificat şi statuat anterior, este folosit şi aici ca punct de referință pentru a determina dacă putem respinge ipoteza nulă. Tabelul pentru valorile critice ale lui *r* (Tabelul 11.4) ilustrează punctul la care semnificația statistică este realizată pentru diferite intensități ale coeficientului de corelație, depinzând de mărimea eşantionului. Aşa cum arată tabelul, pentru un eşantion de 7 indivizi, este nevoie de un coeficient de corelație de 0,754 cu o ipoteză nedirecțională, care ajunge la o semnificație statistică de nivelul de 0,05 (0,874 pentru nivelul 0,01) şi care permite astfel respingerea ipotezei nule. Pentru o ipoteză direcțională ca in exemplul de mai sus, o valoare *r* de 0,874 este semnificativă chiar la nivelul 0,005. Cu o mărime a eşantionului de 400 de cazuri, respingerea ipotezei nule este posibilă cu un coeficient de corelație mult mai slab (0,098 la nivelul 0,05 şi 0,128 la nivelul 0,01). Acesta este un punct important care trebuie reținut.

Ca şi la analiza de asociere, şi la analiza corelației probabilitatea demonstrării semnificației statistice cu ajutorul lui *r* este legată direct de mărimea eşantionului. Într-un eşantion mai mic este mult mai probabil ca întâmplarea (şansa), sub forma erorii de eşantionare, să fie cauza legăturii aparente dintre două variabile, decât într-unul mai mare. Cu un eşantion mare, o legătură aparentă, chiar una care la prima vedere apare a fi slabă, este de departe mai puțin probabilă să fie rezultatul şansei. Un tabel cum este Tabelul 11.4 încorporează acest factor când ia în considerare dimensiunea eşantionului în furnizarea nivelului de semnificație.

Tabel 11.4. Valorile critice ale coeficientului de corelație *r*

	Nivel de semnificație pentru un test direcțional				
	.025	**.005**		**.025**	**.005**
	Nivel de semnificație pentru un test nedirecțional				
n	**.05**	**.01**	***n***	**.05**	**.01**
3	.997	.999	38	.320	.413
4	.950	.990	39	.316	.408
5	.878	.959	40	.312	.403
6	.811	.917	41	.308	.398
7	.754	.874	42	.304	.393
8	.707	.834	43	.301	.389
9	.666	.794	44	.297	.384
10	.632	.765	45	.294	.380

11	.602	.735	46	.291	.374
12	.576	.708	47	.288	.372
13	.553	.684	48	.284	.368
14	.532	.661	49	.281	.364
15	.514	.641	50	.279	.361
16	.497	.623	55	.266	.345
17	.482	.606	60	.254	.330
18	.468	.590	65	.244	.317
19	.456	.575	70	.235	.306
20	.444	.561	75	.227	.296
21	.433	.549	80	.220	.286
22	.423	.537	85	.213	.278
23	.413	.526	90	.207	.270
24	.404	.515	94	.202	.263
25	.396	.505	100	.195	.256
26	.388	.496	125	.170	.230
27	.381	.487	150	.159	.210
28	.374	.478	175	.148	.194
29	.367	.470	200	.138	.181
30	.361	.463	300	.113	.148
31	.355	.456	400	.098	.128
32	.349	.449	500	.088	.115
33	.344	.442	600	.080	.105
34	.339	.436	700	.074	.097
35	.334	.430	800	.070	.091
36	.329	.424	900	.065	.086
37	.325	.418	1000	.062	.081

Substanţialitatea coeficientului de corelaţie

Pentru a interpreta intensitatea unui coeficient de corelaţie, trebuie să luăm în considerare modul în care se utilizează o analiză de corelaţie. Chiar dacă se obţine un coeficient de corelaţie statistic semnificativ, valoarea *r* în sine nu este în mod necesar purtătoare de înţeles şi nu reprezintă neapărat un rezultat substanţial. Pentru eşantioane mici, existenţa unei corelaţii semnificative este un rezultat important, dar pentru eşantioane mari semnificaţia statistică nu este suficientă. De exemplu, cercetările care construiesc instrumente de măsurare sunt preocupate de validitatea şi fiabilitatea lor. Punctul tradiţional de la care se poate accepta un instrument ca prezentând siguranţă, cere adesea un coeficient de corelaţie de peste 0,85, pe când punctul pentru a accepta validitatea instrumentului este frecvent de 0,40 sau chiar mai scăzut. In cercetările sociologice (unde eşantionele sunt de regulă mari) doar o corelaţie de la 0,60 in sus se consideră ca fiind substanţială. Dintr-o perspectivă diferită, identificarea faptului că nu există o asociere (r = 0,00) între două variabile despre care se credea că sunt legate, poate reprezenta cel mai important rezultat al unui studiu.

În interpretarea unui coeficient de corelaţie, mai este important să nu îl tratăm ca şi cum ar fi echivalent cu date de nivel interval sau rapoarte; sau să nu facem judecăţi care dau o astfel de impresie. De exemplu, un coeficient de corelaţie de 0,80 *nu* este de două ori mai puternic ca unul de 0,40. De fapt, 0,80 descrie o asociaţie de patru ori mai puternică ($0,80^2 = 0,64$; $0,40^2 = 0,16$; 0,64 / 0,16 = 4) în ceea ce priveşte cantitatea de variaţie pe care o poate explica din variaţia totală a celeilalte variabile. Trebuie ţinut minte că un coeficient de corelaţie atât de puternic ca 0,80, arată un model (pattern) de la care se vor abate foarte puţine excepţii - adică tuturor valorilor mari ale primei variabile le vor corespunde virtual valori mari şi pentru cea de a doua variabilă, şi viceversa. Un coeficient de corelaţie mai slab (aşa ca 0,40) va avea un procentaj mai mare de cazuri care se opun direcţiei de asociere.

Prezentarea rezultatelor unei analize de corelație

Prezentarea unui coeficient de corelație într-un raport de cercetare este simplă. După calcularea lui *r* conform schemei date în Tabelul 11.3 şi formulei lui de calcul, vom găsi nivelul de probabilitate asociat cu *r* -ul obținut; pe baza dimensiunii eşantionului în tabelul 11.4. Acesta ne va permite să determinăm dacă putem respinge ipoteza nulă sau nu. Cu informația astfel obținută, putem prezenta rezultatele noastre într-o frază.

De exemplu:

> *Între numărul orelor de perfecționare (variabilele independentă) şi numărul eşecurilor în intervenție (variabila dependentă) la un eşantion de 7 asistenți sociali de la un centru de consiliere familială, s-a găsit o corelație r = -0,98. Nivelul de semnificație corespunzător acestei valori este p < 0,01, ceea ce indică prezența unei corelații negative, puternic semnificative din punct de vedere statistic.*

11.3. Corelația cu trei sau mai multe variabile

Legăturile bivariate necesită frecvent explicații în continuare. De exemplu, este improbabil să putem prezice cu siguranță cât timp un client va rămâne în tratament doar pe baza unor informații privind severitatea problemei prezentate de client, motivația lui pentru tratament; sau orice altă variabilă singulară. Perspectiva sistemelor şi alte teorii a cauzelor multiple, susțin că există multe variabile care contribuie împreună şi determină comportamentul uman. Acuratețea predicției poate fi îmbunătățită prin extinderea volumului de date disponibile prin includerea mai multor (decât două) variabile în acelaşi timp într-o analiză. Deciziile privind direcția în care să aibă loc extinderea, sursele suplimentare şi tipurile de date necesare îmbunătățirii explicațiilor şi capacităților noastre predictive, constituie adesea următorul pas critic în procesul de analiză a datelor. În acest text, nu ne vom ocupa cu descrierea testelor care examinează corelațiile între trei sau mai multe variabile (unul binecunoscut este coeficientul de corelație parțială). Pentru dezbaterea noastră, este suficient să spunem că ar trebui să ştim despre existența lor şi despre ceea ce pot ele oferi (sau ce nu pot ele oferi) în plus față de o analiză bivariată, cu utilizarea lui *r*.

11.4. Un exemplu

Descrierea situației care generează un studiu

Radu este un asistent social într-o agenție de servicii familiale. El conduce câteva grupuri de tratament, constituite din adolescente. Recent el şi-a dat seama de variabilitatea mare în participarea verbală printre membrii grupului. Virtual toate fetele ar fi trebuit să participe la discuțiile de grup; dar câteva dintre ele nu au făcut niciodată vreun comentariu nesolicitat. El a simțit că aceste fete aveau un nivel scăzut al dorinței de a se implica în grup. După o perioadă de câteva săptămâni, Radu a întrebat câteva dintre fetele mai timide, de ce sunt aşa de rare comentariile lor voluntare. Dintre cele şapte fete întrebate, cinci au dat un răspuns asemănător: fiecare a fost singurul copil în familie şi au fost învățate de către părinți că nu este datoria lor să inițieze comunicarea. Apoi Radu le-a întrebat pe trei dintre cele mai vorbărețe, care înclinau să domine discuțiile din grup, câți frați şi /sau surori au. Răspunsurile lor au fost şase, şapte şi nouă.

Pe baza incursiunii sale limitate, Radu a început să speculeze o posibilă legătură între o variabilă dependentă, neinițierea unui comentariu în tratamentul de grup şi o variabilă independentă, numărul fraților din familie.

Ipoteza ce urmează a fi testată

Din literatură, Radu a aflat tot ce a putut despre fenomenele ca trăsăturile sociale ale copilului singur, modelele de comunicare dintre frați şi variațiile în participarea verbală în grupurile de adolescenți. Majoritatea literaturii de specialitate părea să-l conducă la concluzia că adolescenții cu mai mulți frați sunt mai dispuşi pentru comentarii voluntare decât aceia cu mai puțini frați. Această presupunere se explica prin experiența în comunicare câştigată prin maturizarea printre frați. Totuşi, altă parte a literaturii de specialitate părea să sugereze o versiune opusă. Ea susținea că acei copii care au în jur adulți cu care să comunice vor căpăta mai multă iscusință verbală şi vor fi mai puțin intimidați de prezența adulților. Astfel

Radu a rămas indecis. Oriunde era vorba de o astfel de temă, în diferitele surse, se sugera că variabila dependentă şi cea independentă par a fi logic legate. Dar în ce direcţie? Radu avea şi el o oarecare experienţă câştigată (desigur neştiinţifică) din propriile observaţii. Până la urmă, s-a decis să realizeze un studiu de cercetare de dimensiuni mici, care să testeze următoarea ipoteză direcţională:

> *Printre adolescentele dintr-un grup de tratament, există o corelaţie pozitivă între numărul comentariilor nesolicitate şi numărul fraţilor din familie.*

O privire asupra metodologiei

Este o procedură obişnuită în agenţia în care lucra Radu de a se înregistra video sesiunile de tratamente de grup şi de a se folosi apoi de către personalul agenţiei pentru supervizare. Astfel Radu nu avea nici o problemă în privinţa accesului la informaţiile care i-ar putea folosi pentru a-şi testa ipoteza. El a primit permisiunea administratorului agenţiei de a folosi benzile video în cercetarea sa.

Pentru operaţionalizare Radu a definit un caz, ca fiind o adolescentă care a participat la cel puţin 75% dintre şedinţele grupului său într-o perioadă de patru luni. Având identificate 35 de fete care îndeplineau acest criteriu, el a revăzut toate videocasetele împreună cu un coleg care a fost interesat de studiul său. Radu şi prietenul său au dezvoltat o definiţie operaţională pentru variabila "comentariu nesolicitat". După aceea, ei au înregistrat numărul comentariilor nesolicitate realizate de fiecare fată în timpul fiecărei şedinţe. Un comentariu nesolicitat a fost judecat ca fiind făcut doar dacă Radu şi colegul său au simţit că a fost conform cu definiţia.

Radu şi colegul său au adunat numărul comentariilor nesolicitate pentru fiecare dintre fete (cazuri) şi apoi l-au împărţit la numărul de şedinţe la care tinerele au participat. Acest număr le-a furnizat media comentariilor nesolicitate pe o şedinţă pentru fiecare caz (variabila dependentă). Din înregistrările făcute de către agenţie, ei au scos datele pentru variabila independentă, numărul fraţilor pentru fiecare caz. Radu a aşezat măsurătorile pentru fiecare dintre cele două variabile într-un tabel asemănător tabelului 11.2.

Rezultatele

Radu a folosit *r* pentru a determina dacă poate să aibă suport statistic pentru ipoteza sa. Folosind formula pentru *r*, el a aflat că valoarea coeficientului de corelaţie dintre numărul comentariilor nesolicitate pe şedinţă şi numărul de fraţi a fost de 0,340. Gândindu-se la ceea ce înseamnă puterea corelaţiei, el a fost oarecum dezamăgit, dar reamintindu-şi că dimensiunea eşantionului său este mare (35 este relativ mare pentru *r*), şi-a dat seama că nu avea nevoie de o corelaţie mare pentru a avea semnificaţie statistică. Când s-a uitat în tabelul pentru *r*, unde dimensiunea eşantionului este controlată (Tabelul 11.4), el a observat că probabilitatea comiterii unei erori de tipul I în respingerea ipotezei nule cu o corelaţie de 0,340 şi un eşantion de 35 era mai mică decât 0,025. (De semnalat că 0,340 se găseşte la dreapta lui 0,334, dar este în stânga lui 0,430).

Interpretarea rezultatelor şi tragerea concluziilor

Radu a ştiut că nivelul de semnificaţie 0,05 este în general acceptat ca suport pentru o legătură dintre două variabile. El ştia de asemenea că în ceea ce priveşte legătura dintre variabila dependentă şi cea independentă, se află pe tărâmul sigur al statisticii. Totuşi este necesar un al doilea pas pentru interpretare, înainte de a revendica suport statistic pentru ipoteza sa. A fost oare corelaţia dintre variabile în *direcţia* în care el a formulat ipoteza, adică este o corelaţie pozitivă? Radu a ţinut minte că o corelaţie pozitivă între două variabile (Figurile 11.1 şi 11.3) înseamnă ca valorile înalte ale unei variabile tind să se afle printre cazurile care au valori înalte ale celeilalte variabile şi viceversa. Aceasta înseamnă că, pentru aceste date, fetele care au avut valori mari pentru variabila "numărul comentariilor nesolicitate" ar fi trebuit să aibă valori mari şi pentru variabila "numărul fraţilor" şi viceversa. Datele lui Radu arată că este aşa. El a concluzionat că are suport statistic pentru ipoteza sa, în direcţia presupusă.

Radu a fost realist în privinţa rezultatelor sale. El a ştiut că pot fi şi alte explicaţii pentru rezultatul său statistic semnificativ, decât existenţa unei legături adevărate între variabile. Cercetarea sa a fost una de dimensiuni mici, el s-a bazat pe un eşantion care îi convenea cel mai mult: de la agenţia sa şi dintre cazurile lui. Multe efecte deformante şi multe alte variabile ar fi posibil să fi afectat rezultatele obţinute de el.

Acestea ar putea să fie legate de factori ca: (1) o anume deformare creată în interiorul eşantionului ca urmare a cazurilor pierdute; (2) o posibilă inabilitate a lui Radu de a fi un bun mediator pentru fetele care nu au fost obişnuite cu situaţiile de grup; (3) o posibilă distorsiune a evenimentelor dată de limitele echipamentelor de înregistrare video folosite. În plus, o corelaţie de 0,340 dintre variabile nu este *puternică* în sens absolut. Lipsa de încredere absolută în rezultatele sale şi mărimea acestora, l-au determinat pe Radu să nu publice încă un articol din studiul său pentru a comunica rezultatele sale şi altora.

11.5. Folosirea SPSS in analizele de corelaţie

Deschideţi fişierul de date conţinut în pachetul SPSS numit „World95". Fişierul conţine mai multe variabile, dintre care ne interesează corelaţia între mortalitatea infantilă – număr de decese la 1000 naşteri vii (*babymort*) şi produsul intern brut pe cap de locuitor (*gdp_cap*), şi corelaţia între mortalitatea infantilă şi procentul femeilor care citesc (*lit_fema*). Pentru aceasta, vom specifica toate cele trei variabile în rubrica *Variables*. Programul va calcula corelaţii între toate cele trei variabile, luate două câte două (chiar dacă pe noi, de pildă, nu ne interesează corelaţia între produsul domestic brut pe cap de locuitor şi procentul femeilor care citesc).

Pentru a calcula coeficientul de corelaţie între două variabile prin programul SPSS, vom urma seria de opţiuni de meniu: ***Analyze – Correlate – Bivariate***. Ultima opţiune deschide o fereastră de dialog conţinând lista variabilelor din fişierul de date (specificate în *Variables*), din care vom alege variabilele între care dorim să se calculeze corelaţia, coeficientul pe care dorim să-l folosim – desigur, în funcţie de tipul de date de care dispunem – (*Pearson, Kendall's tau-b, Spearman*). Pentru fiecare din coeficienţii de corelaţie calculaţi, putem opta să aplicăm un test de semnificaţie a diferenţei faţă de 0 (*Test of Significance*). De asemenea, prin butonul *Options*, putem cere calcularea unor indici sau valori: medii, abateri standard, etc. (*Means and standard deviations, Cross-products deviations and covariances*). Iată ce afişează programul după executarea comenzilor din meniu:

```
CORRELATIONS
  /VARIABLES=babymort gdp_cap lit_fema
  /PRINT=TWOTAIL NOSIG
  /MISSING=PAIRWISE .
```

Correlations

		Infant mortality (deaths per 1000 live births)	Gross domestic product / capita	Females who read (%)
Infant mortality (deaths per 1000 live births)	Pearson Correlation	1	-.640**	-.843**
	Sig. (2-tailed)	.	.000	.000
	N	109	109	85
Gross domestic product / capita	Pearson Correlation	-.640**	1	.429**
	Sig. (2-tailed)	.000	.	.000
	N	109	109	85
Females who read (%)	Pearson Correlation	-.843**	.429**	1
	Sig. (2-tailed)	.000	.000	.
	N	85	85	85

**. Correlation is significant at the 0.01 level (2-tailed).

După ce afişează sintaxa corespunzătoare opţiunilor de meniu executate, programul tipăreşte un tabel în care sunt calculaţi coeficienţii de corelaţie între variabile, luate două câte două. Observăm că valoarea coeficientului de corelaţie între mortalitatea infantilă şi procentul femeilor care citesc este -.843, iar între mortalitatea infantilă şi produsul intern brut pe cap de locuitor este -.64. Sub fiecare dintre coeficienţii calculaţi, sunt trecute rezultatele testelor de semnificaţie (nivelul de încredere e peste 99%

pentru că p=0,000 în fiecare caz) şi aşa cum se specifică în nota **, fiecare dintre coeficienţii calculaţi sunt semnificativ diferiţi de 0 (aşadar, există corelaţie între variabilele respective). Ultima parte a fiecărui rând dă informaţii despre variabile (N reprezentând numărul de cazuri valide folosite la calculul coeficienţilor).

Pe baza rezultatelor obţinute de SPSS, putem concluziona că există o corelaţie negativă puternică între mortaliatea infantilă şi procentul femeilor care citesc, cu alte cuvinte, cu cât procentul populaţiei feminine care citeşte este mai redus, cu atât rata mortalităţii infantile e mai mare. Coeficientul de corelaţie între celelalte două variabile investigate indică de asemenea o corelaţie negativă, mai slabă decât în cazul primelor două variabile, care ne permite să afirmăm că ţările care au un produs domestic brut pe cap de locuitor mai ridicat, înregistrează o rată a mortalităţii infantile mai redusă.

Atenţie: Puteţi folosi corelaţia pentru a identifica intensitatea şi direcţia legăturii dintre două variabile, dar nu şi pentru a stabili o relaţie de cauzalitate. Analiza de corelaţie de mai sus ne arată că, în general, ţările cu un nivel economic mai ridicat şi cu o pondere mai mare de alfabetizare a femeilor, au o mortalitate infantilă mai scăzută, dar nu şi că, de pildă, mortalitatea infantilă e mai ridicată deoarece femeile sunt într-o pondere mai mare analfabete.

Întrebări pentru studiu

1. Care va fi forma unei diagrame de împrăştiere care reflectă o corelaţie negativă perfectă?
2. Ce coeficient de corelaţie sugerează o legătură mai puternică între două variabile, 0,74 sau -0,86?
3. Ce coeficient de corelaţie sugerează că nu există nici o legătură între două variabile de nivel interval?
4. Cum influenţează dimensiunea eşantionului, mărimea coeficientului de corelaţie considerat semnificativ statistic?
5. Cum se poate ca o corelaţie foarte slabă, de 0,10 , să fie o contribuţie valoroasă la îmbunătăţirea cunoştinţelor practice?
6. De ce nu este corect să afirmăm că un coeficientul de corelaţie de 0,84 sugerează o legătură de două ori mai puternică decât un coeficient de corelaţie de 0,42?
7. De ce analizele bivariate, asemenea lui *r*, nu furnizează în mod uzual o explicaţie totală asupra legăturii dintre două variabile?
8. Poate o analiză bivariată de corelaţie să ne spună care variabilă cauzează variaţia în a doua variabilă? Explicaţi.
9. Când poate un coeficient de corelaţie să fie folosit pentru a prezice, cu siguranţă de 100%, valorile unei variabile cu ajutorul valorilor cunoscute pentru cealaltă variabilă?
10. Care este formatul obişnuit pentru prezentarea rezultatelor din analiza de corelaţie?

Exerciţiu pentru acasă 10

Deschideţi fişierul SPSS "World95.sav". Formulaţi o ipoteză de cercetare care să lege fertilitatea populaţiilor ţărilor lumii de diverşi factori cum ar fi: nivelul de dezvoltare economica, nivelul general de educaţie, sau gradul de urbanizare. Cereţi SPSS să calculeze coeficientul de corelaţie Pearson (r) dintre variabila număr mediu de copii la o femeie (fertility) pe de o parte, şi venitul mediu pe cap de locuitor (gdp_cap), procentul de femei alfabetizate (lit_fema) şi procentul populaţiei urbane (urban) pe de altă parte. Care dintre corelaţii este mai puternică şi care este valoarea lui r ? Este ea semnificativă statistic?

ANEXE

Bibliografia completa a cursului

Rotariu T., Bădescu G., Culic I., Mezei E., Mureşan C., 1999, 2006 - *Metode statistice aplicate în ştiinţele sociale*, Editura Polirom, 334 p.

Weinbach R.W., Grinnell R.M. Jr., 1987 - *Statistics for Social Workers*, New York & London

Pah I., 2004 - *Introducere în SPSS*, Editura Eikon, Cluj-Napoca

Atherton C.R., Klemmack D.L., 1991 - *Research Methods in Social Work, An introduction*, University of Alabama

Freedman D., Pisani R., Purves R., 1998 – *Statistics* (3rd ed). Norton & Company Inc., New York

Jaba E., Grama A., 2004 – Analiza statistica cu SPSS sub Windows, Editura Polirom

Norusis, J. Marija, 1992 – *SPSS for Windows. Basw System User's Guide*, SPSS Inc.

Pah I., 2004 - *Tehnici de analiză a datelor cu SPSS*, Presa Universitară Clujeană, Cluj-Napoca

Popa Marian, 2004 – *Statistica psihologica*, Universitatea Bucuresti, http://marianpopa.tripod.com/

Radu I., Micle M., Albu M., Moldovan O., Nemes S., Szamoskozy S., 1993 – *Metodologie psihologica si analiza datelor*, Editura Sincron

Scurtă biografie a titularului de curs

Conf. univ. dr. Cornelia Mureşan este titular la Catedra de Asistenţă socială din cadrul Universităţii Babeş-Bolyai. Domeniile sale de competenţă sunt: metodele cantitative de analiza a datelor in ştiinţele sociale (metode elementare, metode avansate, analiza demografica, metode specifice pentru analiza evenimentelor de viaţă) şi demografia sau studiul populaţiilor (fertilitatea si avortul, familia si copiii, îmbătrânirea si relaţiile dintre generaţii, populaţiile vulnerabile, comportamente demografice). Cornelia Mureşan este membru fondator al *Centrului de Studiere a Populaţiei* si membru in comitetul editorial al *Romanian Journal of Population Studies*. Ea a initiat si coordonat 10 proiecte de cercetare (dintre care 5 internaţionale), a fost membru in colectivele de cercetare in alte 7 contracte naţionale, a publicat ca autor peste 50 de studii ştiinţifice (dintre care 8 in reviste cotate ISI), a publicat 3 cărţi ca autor unic/prim (in domeniul demografiei), 9 cărţi in calitate de coautor (in domeniul metodelor cantitative de analiză şi a bazelor de date), a editat si coordonat 5 volume centrate pe probleme sociale şi de populaţie, a semnat 2 manuale didactice etc. (vezi www.socasis.ubbcluj.ro).